



EDUCACIÓN

SECRETARÍA DE EDUCACIÓN PÚBLICA



TECNOLÓGICO
NACIONAL DE MÉXICO

Tecnológico Nacional de México

Centro Nacional de Investigación
y Desarrollo Tecnológico

Tesis de Doctorado

SEGMENTACIÓN DE HOJAS DE PLANTAS DE JITOMATE,
MEDIANTE APRENDIZAJE PROFUNDO

Presentado por

M.C. Andros Meraz Hernández

como requisito para la obtención del grado de

Doctor en Ciencias de la Computación

Directora de tesis

Dra. Andrea Magadán Salazar

Codirector de tesis

Dr. Jorge Alberto Fuentes Pacheco

Cuernavaca, Morelos, México, octubre de 2026.

 Centro Nacional de Investigación y Desarrollo Tecnológico	ACEPTACIÓN DE IMPRESIÓN DEL DOCUMENTO DE TESIS DOCTORAL		Código: CENIDET-AC-006-D20
			Revisión: 0
	Referencia a la Norma ISO 9001:2008 7.1, 7.2.1, 7.5.1, 7.6, 8.1, 8.2.4		Página 1 de 1

Cuernavaca, Mor., a 10 de septiembre de 2026

DR. CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO
PRESENTE

AT'n: DRA. ANDREA MAGADÁN SALAZAR
PRESIDENTA DEL CLAUSTRO DOCTORAL DEL
DEPARTAMENTO DE CIENCIAS COMPUTACIONALES

Los abajo firmantes, miembros del Comité Tutorial del estudiante **M.C. ANDROS MERAZ HERNÁNDEZ** manifiestan que después de haber revisado el documento de tesis titulado **"Segmentación de Hojas de Plantas de Jitomate, mediante Aprendizaje Profundo"**, realizado bajo la dirección de la **Dra. Andrea Magadán Salazar** y codirigido por el **Dr. Jorge Alberto Fuentes Pacheco**, el trabajo se **ACEPTA** para proceder a su impresión.

ATENTAMENTE

Excelencia en Educación Tecnológica®
"Conocimiento y Tecnología al Servicio de México"


DRA. ANDREA MAGADÁN SALAZAR
TECNM/CENIDET


DR. JORGE ALBERTO FUENTES PACHECO
TECNM/CENIDET


DR. RAÚL PINTO ELÍAS
TECNM/CENIDET


DR. NIMROD GONZÁLEZ FRANCO
TECNM/CENIDET


DRA. MARÍA YASMÍN HERNÁNDEZ PÉREZ
TECNM/CENIDET


DR. GINER ALOR HERNÁNDEZ
TECNM / IT. DE ORIZABA

c.c.p: Verónica Sotelo Boyas / Jefa del Departamento de Servicios Escolares
c.c.p: Nôe Alejandro Castro Sánchez / Jefe del Departamento de Ciencias Computacionales
c.c.p: Expediente



Centro Nacional de Investigación y Desarrollo Tecnológico
Subdirección Académica

Cuernavaca Mor, 17/septiembre/2026

Oficio No. SAC/308/2026

Asunto: Autorización de impresión de tesis

ANDROS MERAZ HERNÁNDEZ
CANDIDATO AL GRADO DE DOCTOR EN CIENCIAS
DE LA COMPUTACIÓN
P R E S E N T E

Por este conducto, tengo el agrado de comunicarle que el Comité Tutorial asignado a su trabajo de tesis titulado **"Segmentación de Hojas de Plantas de Jitomate, mediante Aprendizaje Profundo"**, ha informado a esta Subdirección Académica, que están de acuerdo con el trabajo presentado. Por lo anterior, se le autoriza a que proceda con la impresión definitiva de su trabajo de tesis.

Esperando que el logro del mismo sea acorde con sus aspiraciones profesionales, reciba un cordial saludo.

ATENTAMENTE

Excelencia en Educación Tecnológica
"Conocimiento y Tecnología al servicio de México"

CARLOS MANUEL ASTORGA ZARAGOZA
SUBDIRECTOR ACADÉMICO



c.c.p. Departamento de Ciencias Computacionales
Departamento de Servicios Escolares

CMAZ/lmz



2026
año de
Margarita
Maza

cenidet
Centro Nacional de Investigación
y Desarrollo Tecnológico



Interior Internado Palmira S/N, Col. Palmira,
C.P. 62490, Cuernavaca, Morelos Tel. 01 (777) 3627770, ext. 4106,
e-mail: acad_cenidet@tecnm.mx | cenidet.tecnm.mx

DEDICATORIAS

A mi madre † que fue mi gran pilar en vida.

A mi padre por los grandes sacrificios hechos, compañía y apoyo incondicional brindados en todo momento. Gracias por ser mi otro pilar de vida.

AGRADECIMIENTOS

Doy gracias a Dios por haberme brindado fortaleza en tiempos complicados y permitirme culminar este logro académico importante en mi vida ¡Gracias Dios!

A mi familia y amistades cercanas de mi madre por su acompañamiento y apoyo brindado.

A mis principales guías de esta investigación Dra. Andrea Magadán Salazar y Dr. Jorge Fuentes Pacheco por su atención y tiempo dedicado en asesorías y recomendaciones. Expreso mi sincero agradecimiento por haberme dado nuevamente la oportunidad de formar parte del equipo de IA en CENIDET.

A mi comité revisor: Dra. María Yasmín Hernández Pérez, Dr. Raúl Pinto Elías, Dr. Nimrod González Franco y, Dr. Giner Alor Hernández por sus correspondientes retroalimentaciones hacia la mejora de este trabajo de investigación y crecimiento profesional.

Agradezco especialmente el apoyo económico brindado por parte de la Secretaría de Ciencia Humanidades Tecnología e Innovación (SECIHTI) durante la realización de mis estudios de posgrado.

Al Tecnológico Nacional de México / Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), por dejarme las puertas abiertas para continuar con mi formación académica y profesional.

A mis estimados profesores del Tecnológico Nacional de México / campus Zacatepec: Dra. Josefina Sámano Galindo, Dra. Ana Celia Campos Hernández, Dra. Teresa Vichis Rojas, M. en M. Andrés Hernández Peralta, M. en C. Mario Humberto Tiburcio Zúñiga y, M. en C. José Antonio Vázquez Santa Anna por el apoyo y conocimientos brindados.

Finalmente, un cálido agradecimiento, a todos mis compañeros(as) y amigos(as) que están y estuvieron presentes e hicieron especial esta etapa a lo largo de estos cuatro años.

Contenido

CAPÍTULO 1 INTRODUCCIÓN	1
1.1 Introducción	1
1.2 Descripción del problema	2
1.3 Complejidad del problema	2
1.4 Preguntas de investigación	3
1.5 Objetivos	4
1.6 Alcances	4
1.7 Limitaciones	4
1.8 Justificación	5
1.9 Hipótesis	5
1.10 Método de solución	5
1.11 Contenido	6
CAPÍTULO 2 ANTECEDENTES	8
2.1 Estudios previos	8
2.2 Estado del arte	8
2.3 Conclusiones del estado del arte	15
CAPÍTULO 3 MARCO TEÓRICO	18
3.1 Cultivo bajo estudio	18
3.2 Tópicos computacionales	21
3.3 Redes neuronales especializadas en detección	26
3.4 Redes especializadas en clasificación	30
3.5 Redes enfocadas en segmentación	34
3.6 Configuraciones específicas	37
3.7 Funciones de regresión para el cálculo de instancias delimitadoras	40
3.8 Métricas de evaluación	41
CAPÍTULO 4 BANCO DE IMÁGENES	45
4.1 Bancos de imágenes provenientes de <i>RoboFlow</i>	45
4.2 Agrupación de imágenes web	50
4.3 Conjunto para la etapa de pruebas	52
4.4 Conjunto de imágenes de otros cultivos	53
CAPÍTULO 5 MÉTODO DE SOLUCIÓN	56

5.1 Etapas del método propuesto	56
5.2 Modificaciones en la arquitectura de la red <i>Yolo11n</i>	58
5.3 Sustitución de la función <i>CloU</i> de la red <i>Yolo11n</i>	63
CAPÍTULO 6 ANÁLISIS Y DISEÑO DEL SISTEMA	66
6.1 Requerimientos funcionales.....	66
6.2 Casos de uso	66
6.2 Maquetado del sistema móvil.....	69
CAPÍTULO 7 EXPERIMENTACIÓN Y RESULTADOS	71
7.1 Experimentación	71
7.2 Estudio de ablaciones con la arquitectura de <i>Yolov11n</i>	74
7.3 Validación cruzada.....	78
7.4 Comparación de <i>YoloLeaf</i> contra <i>Yolov12n</i> y <i>Yolov11n</i>	80
7.5 Análisis cuantitativo por casos, <i>YoloLeaf</i> contra <i>Yolov12n</i> y <i>Yolov11n</i>	83
7.6 Pruebas de localización con hojas de jitomate	86
7.7 Comparación cualitativa con el estado del arte	97
7.8 Evaluación cualitativa con otros tipos de hojas del jitomate	98
7.9 Evaluación con otros tipos de cultivos	99
7.10 Comparación cuantitativa: <i>Yolov12n</i> , <i>Yolov11n</i> vs <i>YoloLeaf</i>	110
7.11 Resultados de la segmentación de hojas	112
7.12 Segmentación de hojas con otros tipos de cultivos.....	118
CAPÍTULO 8 CONCLUSIONES.....	121
8.1 Conclusiones generales	121
8.2 Actividades Académicas	124
8.3 Trabajo a futuro.....	129
Referencias	131

ÍNDICE DE FIGURAS

Figura 1.1. Ejemplo de alteraciones en la lámina foliar	2
Figura 1.2. Desafíos y perspectiva de captura.	3
Figura 1.3. Etapas del método propuesto.	5
Figura 2.1. Conteo de las principales editoriales.	9
Figura 2.2. Evidencias de los bancos de imágenes.....	17
Figura 3.1. Tipos de tomate según su hábito de crecimiento (López. et. al, 2004).	19
Figura 3.2. Diferentes especies naturales del jitomate (Zsögön et. at., 2017).	19
Figura 3.3. Mancha bacteriana (MSU, s.f.).	20
Figura 3.4. Hoja con moho (MSU, s.f.).	21
Figura 3.5. Enfermedad del virus de la hoja amarilla (MSU, s.f.).	21
Figura 3.6. Representación del aprendizaje profundo utilizando un modelo de clasificación de imágenes (Chollet, F, 2018).	23
Figura 3.7. Representación gráfica del operador convolucional (Hüseyin et al., 2022).	24
Figura 3.8. Representación gráfica de una operación de Max Pooling (Parreño, 2018). ...	25
Figura 3.9. Arquitectura YOLO (Redmon et al., 2016).	26
Figura 3.10. Módulos clave en la arquitectura <i>Yolov11</i> (Rahima Khanam & Muhammad Hussain, 2024).	27
Figura 3.11. Área de atención (Tian et al., 2025).	29
Figura 3.12. Módulo residual R-ELAN (Tian et al., 2025).	30
Figura 3.13. Representaciones de atención (Dosovitskiy et al., 2021).	30
Figura 3.14. Transformador visual (ViT) (Dosovitskiy et al., 2021).	32
Figura 3.15. Arquitectura MobileViT (Sachin & Mohammad, 2022).	33
Figura 3.16. Multiconexión entre píxeles en el bloque MobileViT (Sachin & Mohammad, 2022).	34
Figura 3.17. Esquema del modelo SAM (Chaoning et al., 2023).	34
Figura 3.18. Representación de dos tipos de destilación del conocimiento acoplado del modelo SAM (Chaoning et al., 2023).	35
Figura 3.19. Destilación desacoplada (Chaoning et al., 2023).	35
Figura 3.20. Predicción de una máscara de píxeles, delimitando al objeto con un BB (Chaoning et al., 2023).	36
Figura 3.21. Topología de SAM2 (Ravi et al., 2024).	37
Figura 3.22. Comparativa de operaciones convolucionales (Hüseyin et al., 2022).	37
Figura 3.23. Esquema general de CBAM (Woo et al., 2018).	38
Figura 3.24. Diagrama del módulo de atención de canal (Woo et al., 2018).	39
Figura 3.25. Diagrama del módulo de atención espacial (Woo et al., 2018).	39
Figura 3.26. The mapping of outlier degree (Tong et al., 2023).	41
Figura 3.27. Comparación de máscaras de píxeles.	42
Figura 3.28. Comparación de cajas delimitadoras.	43
Figura 4.1. Capturas digitales del banco de imágenes A4 (Roboflow, s.f).	45
Figura 4.2. Capturas digitales del banco de imágenes T7 (Roboflow, s.f).	46
Figura 4.3. Capturas digitales del banco de imágenes PCVP (Roboflow, s.f).	46
Figura 4.4. Muestras digitales del banco de imágenes Tomato Test (Roboflow, s.f).	47

Figura 4.5. Tomas de la agrupación de imágenes Answer Computer Vision Project (Roboflow, s.f).....	47
Figura 4.6. Fotos de la agrupación de imágenes Detecting Calciumn (Roboflow, s.f).	48
Figura 4.7. Tomas del conjunto NAIRS Computer Vision Project (Roboflow, s.f).	48
Figura 4.8. Imágenes del conjunto Tomato Computer Vision Project (Roboflow, s.f).....	49
Figura 4.9. Imágenes del conjunto Tomato Data Leaf 940 (Roboflow, s.f).....	49
Figura 4.10. Imágenes del conjunto TLD (Roboflow, s.f).....	50
Figura 4.11. Collage de hojas de la planta del jitomate.	50
Figura 4.12. Etiquetado de múltiples hojas.	51
Figura 4.13. Collage del conjunto de pruebas.	53
Figura 4.14. Ejemplo de cultivos seleccionados (Ilyas et al, 2025).....	53
Figura 5.1. Flujo de trabajo de la propuesta de solución.	56
Figura 5.2. Backbone original y modificado; remplazo de la función de regresión.....	57
Figura 5.3. La figura esquematiza: a) módulo original MV2Block y b) inclusión de CBAM en MV2Block.	60
Figura 5.4. La figura ilustra: a) módulo original MobileViTBlock y b) remoción de capas Conv (SC) en el módulo MobileViTBlock.	61
Figura 5.5. Representación gráfica de las variables.....	63
Figura 5.6. Aspectos y solapamientos de instancias delimitadoras.	64
Figura 5.7. Esquematización gráfica del AB y BB.....	64
Figura 5.8. Comparación del comportamiento monótono y no monótono.	65
Figura 6.1. Diagrama de casos de uso.	66
Figura 6.2. Diagrama de flujo del proceso del proceso de localización.....	67
Figura 6.3. Diagrama de flujo del proceso de segmentación.	68
Figura 6.4. Maquetado de la aplicación móvil.	69
Figura 7.1. Propiedades de las instancias del banco de imágenes propuesto.	72
Figura 7.2. Curvas de aprendizaje de YoloLeaf.	76
Figura 7.3. Curvas de evaluación de YoloLeaf.....	77
Figura 7.4. Curva F1-Confidence.....	78
Figura 7.5. Distribución de instancias.	83
Figura 7.6. Análisis de las métricas <i>Recall</i> y <i>Precision</i>	85
Figura 7.7. Tipos de hojas de la planta del jitomate (Rahmadian et al., 2024).	99

ÍNDICE DE TABLAS

Tabla 2.1. Síntesis del estado del arte relacionado a la localización y/o segmentación de cultivos.	9
Tabla 3.1. Técnicas de segmentación (Xicohtencatl, 2018).....	22
Tabla 3.2. Comparación de parámetros y rapidez de inferencia (Chaoning et al., 2023). .	36
Tabla 3.3. Matriz de confusión.	42
Tabla 3.4. Métricas de evaluación.	43
Tabla 4.1. Composición del banco de imágenes propuesto.....	52
Tabla 4.2. Distribución de ejemplares del cacahuete.....	54
Tabla 4.3. Distribución de ejemplares del frijol.	54
Tabla 4.4. Distribución de ejemplares del frijol rojo.	55
Tabla 5.1. Componentes originales en Yolo11n.	58
Tabla 5.2. Componentes originales en MobileViT.	58
Tabla 5.3. Configuraciones en Yolo11n.....	58
Tabla 5.4. Código de colores.....	59
Tabla 5.5. Módulos del backbone en la arquitectura propuesta.	62
Tabla 5.6. Módulos originales del backbone en la arquitectura Yolov11n.	63
Tabla 6.1. Explicación de cada etapa del proceso de detección.....	68
Tabla 6.2. Explicación de cada etapa del proceso de segmentación.	69
Tabla 7.1. Distribución de los ejemplares.....	72
Tabla 7.2. Hiperparámetros genéricos de la familia Yolo.	73
Tabla 7.3. Casos de experimentación modificando Yolo11n.	75
Tabla 7.4. Resultados alcanzados con las ocho configuraciones del modelo.	75
Tabla 7.5. Configuración de la validación cruzada realizada.	79
Tabla 7.6. Equiparación de resultados.	79
Tabla 7.7. Análisis estadístico.	79
Tabla 7.8. Comparación de características y métricas I.	80
Tabla 7.9. Comparación de características y métricas II.	80
Tabla 7.10. Comparación del costo beneficio con tendencia descendente.	81
Tabla 7.11. Comparación del costo beneficio con tendencia ascendente.	81
Tabla 7.12. Comparación de matrices de confusión.	82
Tabla 7.13. Análisis del desempeño de los tres modelos.....	83
Tabla 7.14. Conteo del desempeño superior.....	86
Tabla 7.15. Evaluación cualitativa y cuantitativa de localización (oclusiones).....	86
Tabla 7.16. Evaluación cualitativa y cuantitativa de localización (cambio en la forma).....	87
Tabla 7.17. Evaluación cualitativa y cuantitativa de localización (daño por plaga).	88
Tabla 7.18. Evaluación cualitativa y cuantitativa de localización (agrupación de hojas sanas).....	88
Tabla 7.19. Evaluación cualitativa y cuantitativa de localización (agrupación de hojas anómalas).....	89
Tabla 7.20. Evaluación cualitativa y cuantitativa de localización (iluminación irregular I)..	90
Tabla 7.21. Evaluación cualitativa y cuantitativa de localización (Iluminación irregular II).	91
Tabla 7.22. Evaluación cualitativa y cuantitativa de localización (presencia de jitomate I).	92

Tabla 7.23. Evaluación cualitativa y cuantitativa de localización (presencia de jitomate II).	93
Tabla 7.24. Evaluación cualitativa y cuantitativa de localización (presencia de insecto I).	94
Tabla 7.25. Evaluación cualitativa y cuantitativa de localización (presencia de insecto II).	95
Tabla 7.26. Evaluación cualitativa y cuantitativa de localización (objetos verdes).	96
Tabla 7.27. Evaluación cualitativa y cuantitativa de localización (dibujos o impresiones).	97
Tabla 7.28. Resultados del conjunto de prueba.	97
Tabla 7.29. Comparación de inferencias YOLO-LeafNet contra YoloLeaf.	98
Tabla 7.30. Características cualitativas del tipo de hoja de la planta del jitomate (Rahmadian et al., 2024).	99
Tabla 7.31. Evaluación cuantitativa del cacahuete.	100
Tabla 7.32. Evaluación cualitativa del cacahuete.	101
Tabla 7.33. Evaluación cuantitativa del frijol.	103
Tabla 7.34. Evaluación cualitativa del frijol.	104
Tabla 7.35. Evaluación cuantitativa del frijol rojo.	106
Tabla 7.36. Evaluación cualitativa del frijol rojo.	108
Tabla 7.37. Síntesis comparativa de Yolov12n, Yolov11n y YoloLeaf.	110
Tabla 7.38. Evaluación cualitativa.	111
Tabla 7.39. Evaluación cuantitativa del desempeño.	112
Tabla 7.40. Análisis de resultados de segmentación (oclusión).	112
Tabla 7.41. Análisis de resultados de segmentación (cambio de forma).	113
Tabla 7.42. Análisis de resultados de segmentación (daño por plaga).	113
Tabla 7.43. Análisis de resultados de segmentación (cambio en aspecto).	114
Tabla 7.44. Análisis de resultados de segmentación (iluminación irregular).	114
Tabla 7.45. Análisis de resultados de segmentación (solapamientos).	115
Tabla 7.46. Análisis de resultados de segmentación (oclusión).	115
Tabla 7.47. Análisis de resultados de segmentación (cambio en la tonalidad).	116
Tabla 7.48. Análisis de resultados de segmentación (proyección de sombras).	116
Tabla 7.49. Análisis de resultados de segmentación (iluminación a contraluz).	117
Tabla 7.50. Análisis de resultados de segmentación (cambios de coloración).	117
Tabla 7.51. Análisis de resultados de segmentación (hoja a través de objeto transparente).	118
Tabla 7.52. Análisis de resultados de segmentación (lámina foliar no anómala).	118
Tabla 7.53. Láminas foliares segmentadas.	119
Tabla 7.54. Evaluación cualitativa y cuantitativa del funcionamiento.	120
Tabla 8.1. Cumplimiento del objetivo general.	122
Tabla 8.2. Cumplimiento de los objetivos específicos.	122
Tabla 8.3. Cumplimiento de los alcances.	123

ÍNDICE DE ECUACIONES

Ecuación 3.1.....	22
Ecuación 3.2.....	22
Ecuación 3.3.....	24
Ecuación 3.4.....	25
Ecuación 3.5.....	25
Ecuación 3.6.....	26
Ecuación 3.7.....	31
Ecuación 3.8.....	31
Ecuación 3.9.....	31
Ecuación 3.10	31
Ecuación 3.11	38
Ecuación 3.12	38
Ecuación 3.13	38
Ecuación 3.14	38
Ecuación 3.15	39
Ecuación 3.16	39
Ecuación 3.17	40
Ecuación 3.18	40
Ecuación 3.19	40
Ecuación 3.20	40
Ecuación 3.21	40
Ecuación 3.22	41
Ecuación 3.23	41
Ecuación 5.1.....	62
Ecuación 5.2.....	63
Ecuación 7.1.....	80

ACRÓNIMOS

Abreviatura	Significado
AB(es)	Anchor Box(es)
AI o IA	Artificial Intelligence o Inteligencia Artificial
ANN o RNA	Artificial Neuronal Network o Red Neuronal Artificial
AV o VA	Artificial Vision o Visión Artificial
BB(es)	Bounding Box(es)
CIoU	Complete Intersection over Union
CBAM	Convolution Block Attention Module
CNN	Convolutional Neuronal Network
CVAT	Computer Vision Annotation Tool
DL	Deep Learning
DWC	Deep Wise Convolution
FP	False Positive
FN	False Negative
FLOPS	Floating Point Operations Per Second
GPU	Graphics Processing Unit
GT	Ground Truth
HSI	Hue, Saturation, Intensity
IDE	Integrated Development Environment
IoU	Intersection over Union
ML	Machine Learning
MLP	Multi Layer Perceptron
NDK	Native Developer Kit
PWC	Point Wise Convolution
Q,K,V	Query, Key, Value
RAM	Random Access Memory
RGB	Red Green Blue
ROI	Region of Interest
SAM	Segment Anything Model
SGD	Stochastic Gradient Descent
SiLU	Sigmoid Linear Unit
TP	True Positive
TN	True Negative
ViT	Vision Transformer
YOLO	You Only Look Once
WIoU	Wise Intersection over Union

RESUMEN

En los últimos años, es común encontrar en la literatura trabajos relacionados con la agricultura de precisión que incorporan el uso de redes neuronales profundas como parte parcial o total de una propuesta de solución.

La segmentación de las hojas en las imágenes constituye una de las etapas más importantes en los sistemas de diagnóstico automático como enfermedades foliares, ya que permite identificar con precisión las regiones de interés. Su adecuada implementación impacta directamente en la precisión e interpretabilidad de los modelos de aprendizaje profundo.

En esta tesis se presenta el desarrollo de un método para la localización y segmentación de hojas de la planta del jitomate a través de imágenes capturadas a campo abierto; esto implica considerar una serie de factores que afectan el desempeño del sistemas tales como: condiciones climáticas (cambios de iluminación), estado de salud de la planta (falta de nutrientes, patologías y/o estrés) que alteran el aspecto de la hoja significativamente, la presencia de objetos que asemejen la forma de una hoja, dibujos e impresiones, la perspectiva de captura, etc.

El modelo protagonista denominado “*YoloLeaf*” es una versión mejorada de la red *Yolov11n* para inferir la localización de láminas foliares en imágenes capturadas bajo diferentes condiciones naturales adversas, detectando el mayor número de hojas existentes. Las modificaciones realizadas son: un módulo convolucional en dos etapas y, dos módulos de transformadores de visión que en conjunto favorecen a una mejor representación de los mapas de características. Además, se remueve la función *CIoU* por *WIoUv3* para reducir problemas al estimar la ubicación de las hojas de la planta.

En la etapa experimental el modelo se evaluó en 10 escenarios representativos de entornos al aire libre usando el conjunto de validación. *YoloLeaf* registró una mayor recuperación (*Recall*) de hojas existentes en cinco casos de estudio: “frutos rojos” (82.00%), “variación de iluminación” con (82.60%), “presencia de flores” (83.70%), “cambios en las escalas” con (85.60%) y, “punto focal hacia la hoja” (88.60%). Adicionalmente, se registraron menos casos de falsos positivos (*Precision*) en “presencia de flores” (89.00%), “anomalías de coloración en hojas” (86.20%) y “frutos verdes” (78.30%).

Finalmente, se evaluó la generalización de *YoloLeaf* con ejemplares donde figuran cultivos de cacahuete, frijol y frijol rojo (por separado), capturadas con tres ángulos de visión (0°, 45° y 90°) durante las primeras ocho semanas en zonas de cultivo al exterior. El modelo *YoloLeaf* registró una buena inferencia al obtener porcentajes de: 92.00% de *Precision*, 88.70% de *Recall*, 93.70% de *mAP@50* y 86.70% de *mAP@50-95* con imágenes del frijol rojo capturadas a 90° en la cuarta semana de desarrollo.

Abstract

In recent years, deep learning neural networks have emerged as a prominent approach in precision agriculture, being increasingly integrated into both specific components and end-to-end solutions for agricultural applications.

Leaf segmentation is a critical stage in automated diagnostic systems, particularly for the detection and assessment of foliar diseases, as it enables the precise identification and delineation of regions of interest. Accurate leaf segmentation directly affects the performance and interpretability of deep learning models.

This study presents a method for the localization and segmentation of tomato leaves in images acquired under outdoor conditions. The proposed approach addresses several factors that can affect the performance of automated vision systems, including variations in illumination caused by changing weather conditions, plant health status such as nutrient deficiencies, diseases, and stress, the presence of leaf-like objects, patterns, and printed textures, as well as variations in the image acquisition angle. These factors can substantially alter leaf appearance and increase the complexity of accurate leaf localization and segmentation.

The proposed model, YoloLeaf, is an enhanced version of the *Yolov11n* architecture designed to maximize leaf location in images acquired under diverse and challenging natural conditions. The modifications include the incorporation of a Two-Stage Convolutional module and two Vision Transformer modules, which jointly contribute to a more effective representation of the feature maps. In addition, the original CloU loss function is replaced with WIoUv3 in order to mitigate localization errors during the calculation of the spatial position of plant leaves.

YoloLeaf was evaluated across 10 scenarios representative of outdoor environments using the validation dataset. YoloLeaf achieved higher recall for existing leaves in five case studies: “ripening of the tomato fruit” (82.00%), “variation in lighting” (82.60%), “presence of flowers” (83.70%), “changes in scale” (85.60%) and “focusing on the leaf” (88.60%). In addition, fewer false positives (Precision) were recorded for “presence of flowers” (89.00%), “leaf coloration anomalies” (86.20%) and “unripening of the tomato fruit” (78.30%).

In addition, the inference of proposed model was evaluated using images showing peanut, bean and red bean crops (separately), captured from three viewing angles (0°, 45° and 90°) during the first eight weeks in outdoor growing areas. The YoloLeaf model demonstrated good performance, achieving the following percentages: 92.00% Precision, 88.70% Recall, 93.70% mAP@50 and 86.70% mAP@50-95 with images of red beans captured at 90° during the fourth week of growth.

CAPÍTULO 1

INTRODUCCIÓN

En el presente capítulo se expone el análisis del problema, los objetivos, alcances y limitaciones, la justificación y el método de solución propuesto, que constituyen la base del presente trabajo de tesis.

1.1 Introducción

El jitomate (*Solanum lycopersicum* L., antes *Lycopersicon esculentum* Mill) es una de las hortalizas de mayor relevancia agroeconómica global. México se posiciona como el principal proveedor mundial de este cultivo, concentrando el 25.11% del valor de las exportaciones internacionales y abasteciendo mercados críticos como Estados Unidos y Canadá (SAGARPA, 2017). A pesar de las proyecciones de crecimiento que estiman un incremento en la producción nacional de hasta 7.56 millones de toneladas para el año 2030 (SAGARPA, 2017), la intensificación del cultivo enfrenta serios desafíos fitosanitarios y climáticos para su desarrollo (López et al., 2004).

Tradicionalmente, la inspección y el monitoreo de los cultivos dependen de la evaluación visual humana, un método subjetivo, propenso a errores y limitado para cubrir extensiones a gran escala. Ante esta problemática, la comunidad agrícola ha manifestado la necesidad urgente de mejorar los procesos de monitorización para detectar de manera temprana el estrés biótico y abiótico, garantizando así la calidad exigida por los mercados internacionales (González-Pérez et al., 2022). Como respuesta tecnológica a esta demanda, los Sistemas de Visión Artificial (SVA) han emergido en la agricultura de precisión como herramientas esenciales para la clasificación de enfermedades, identificación de deficiencias nutricionales y detección de plagas (Tian et al., 2020). No obstante, la transición de estos sistemas de entornos controlados a escenarios agrícolas reales en campo abierto representa un desafío técnico complejo. El desempeño de los SVA en ambientes exteriores se ve afectado de forma negativa por factores no controlados, tales como las variaciones dinámicas de la iluminación natural, la oclusión foliar y las condiciones climáticas cambiantes (Vázquez-Arellano et al., 2018; Meshram et al., 2021).

Para mitigar estas limitantes en exteriores, los SVA ejecutan múltiples procesos de inferencia robustos. Entre estos, destacan dos etapas fundamentales: la localización de los objetos de interés mediante cajas delimitadoras (*Bounding Boxes*, BBs) que definen las coordenadas espaciales del objetivo, y la segmentación semántica o por instancias, encargada de delimitar de manera precisa y a nivel de píxel las regiones previamente localizadas.

1.2 Descripción del problema

Pese a los avances en el desarrollo de nuevas redes neuronales convolucionales (Rahima Khanam & Muhammad Hussain, 2024), (Tian *et al.*, 2025) y la llegada de los transformadores visuales (Dosovitskiy *et al.*, 2021), la localización de hojas en imágenes capturadas bajo condiciones naturales adversas continúa siendo un problema desafiante que impacta negativamente en la inferencia del modelo, comprometiendo la funcionalidad de un SVA.

La investigación se enfoca en localizar todas las hojas de la planta de jitomate en imágenes tomadas en condiciones de campo adversas presentes en la naturaleza.

1.3 Complejidad del problema

La localización espacial de las hojas en una imagen adquirida en escenarios naturales, a partir de su inspección visual, es una tarea compleja. La dificultad aumenta debido a que la hoja es un organismo vivo que cambia debido a que las condiciones del entorno afectando significativamente su aspecto. Lo que provoca que las láminas foliares presenten alteraciones fenotípicas que se derivan en perturbaciones bióticas y abióticas tales como las que se ilustran en la Figura 1.1, a) estrés hídrico, b) partes necróticas, c) cambio de color por avance de enfermedad, d) falta de nutrientes, e) presencia de plagas como el minador, f) infestación de la mosca blanca, g) invasión de arañas y h) presencia de hongo.

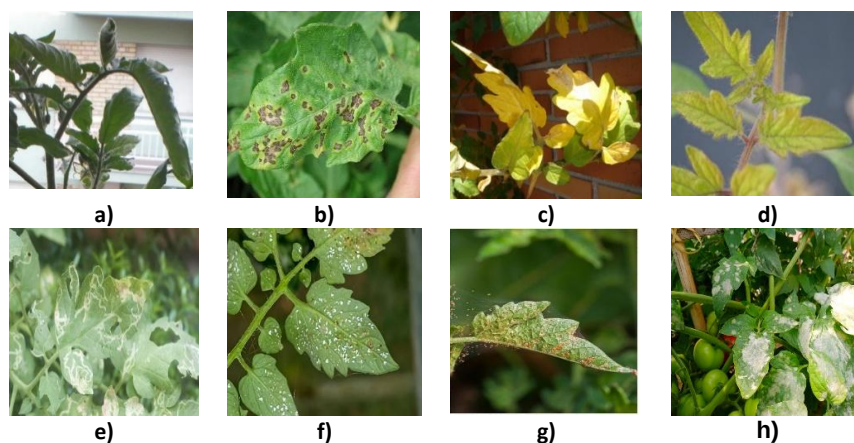


Figura 1.1. Ejemplo de alteraciones en la lámina foliar.

En la Figura 1.1., se ilustra en a) un fenómeno de ambigüedad morfológica donde el enrollamiento de la lámina foliar induce a una distorsión susceptible de ser confundida con un gusano. En b) se ilustra el desafío que representa el mimetismo con el entorno o fondo de la imagen, donde las características visuales son semejantes al objeto bajo estudio. En c) y d) el color amarillo genera una alta similitud cromática con las flores del cultivo. En e) las oclusiones mutuas debido a los solapamientos y el deterioro de la condición de salud de la planta restringe la delimitación precisa de las fronteras del objeto. Además, el campo visual de la cámara pudiera no ser el adecuado debido al ángulo o distancia de captura, ocasionando distorsiones como la visibilidad parcial o nula del área de interés, comprometiendo el desempeño del modelo.

El entorno de captura presenta otros elementos que introducen ruido visual debido a su similitud con el objeto de estudio. La Figura 1.2, ilustra algunos de estos desafíos los cuales abarcan: a) áreas de vegetación circundante, b) gusanos y flores del cultivo), c) frutos no maduros, d) artefactos o herramientas de cultivo como bolsas y macetas. También las hojas presentan una variabilidad adicional, ya que mientras una hoja sana exhibe una coloración uniforme cuando esta está sana (tono verde), el tejido afectado por alguna patología manifiesta dos o más coloraciones en tonalidades verdes, amarillas y cafés.



Figura 1.2. Desafíos y perspectiva de captura.

Adicionalmente, un SVA también se ve afectado por factores tales como las sombras, la incidencia de la luz sobre las hojas que provoca cambios en la tonalidad y la pérdida de características esenciales. Además, la variación de la escala del objetivo dificulta obtener detalles, particularmente en zonas de dimensiones reducidas como las nervaduras secundarias. Finalmente, el ángulo y la perspectiva de captura que no son favorables introducen distorsiones proyectivas alterando, en la imagen, la morfología original del objeto (por ejemplo, alargamientos o acortamientos) originando formas inexistentes en la naturaleza.

1.4 Preguntas de investigación

A partir de la problemática planteada y de la necesidad de desarrollar métodos automatizados que permitan mejorar el funcionamiento de los SVA a campo abierto mediante el empleo de redes neuronales artificiales profundas, se establecen las siguientes preguntas de investigación. Estas interrogantes se derivan del problema de investigación y permiten delimitar los aspectos relacionados con la modificación propuesta, implementación y evaluación del modelo desarrollado:

- ¿Qué arquitecturas de aprendizaje profundo son empleados en la detección y/o segmentación de hojas?
- ¿Cuáles son las modificaciones propuestas que mejoran el desempeño respecto al modelo base?
- ¿Cuáles son los efectos de incorporar módulos de atención basados en transformadores de visión, respecto al desempeño del modelo?

- ¿Cuáles son los beneficios de sustitución de la función de regresión original por *WloUv3* sobre el funcionamiento del modelo?
- ¿Cuál es la razón de emplear dos redes neuronales secuencialmente para abordar el problema de la segmentación del área foliar?
- ¿La incorporación de los módulos de atención y la función *WloUv3* conjuntamente mejora la localización de hojas y, al mismo tiempo reduce el costo computacional con respecto al modelo base?
- ¿La generalización del modelo propuesto localiza otro tipo de hojas que se asemejen con respecto al cultivo bajo estudio?
- ¿El modelo desarrollado puede ser implementado en un dispositivo con recursos limitados manteniendo un desempeño adecuado?

1.5 Objetivos

a) Objetivo general

Modificar la arquitectura de una red neuronal para localizar y segmentar las hojas de las plantas de jitomate en imágenes capturadas a campo abierto.

b) Objetivos específicos

- Analizar el estado del arte relacionado con el tema de detección de hojas y frutos en entornos agrícolas naturales, mediante aprendizaje profundo, con el propósito de identificar las metodologías, el desempeño y limitaciones de las actuales arquitecturas de redes.
- Construir un conjunto de imágenes de plantas de jitomate etiquetadas mediante un protocolo de anotación para generar el *ground truth* requerido en las etapas de entrenamiento y validación de los modelos de redes (original y propuesto).
- Desarrollar modificaciones en la arquitectura seleccionada para mejorar la capacidad de generalización del modelo original en entornos complejos del cultivo de las plantas de jitomate.
- Validar de forma cuantitativa y cualitativa el desempeño de la red propuesta, empleando métricas clásicas de evaluación para contrastar los resultados frente al modelo base y a soluciones competitivas del estado del arte.

1.6 Alcances

- Las imágenes contienen hojas de plantas de jitomate, en diversas etapas de desarrollo (crecimiento, floración y maduración) capturadas en condiciones externas de campo.
- El modelo localiza hojas de jitomate del tipo “regular y papa”.
- El modelo principal (contribución) sólo cuantifica el número de hojas localizadas y delimita la región de interés para cada lámina foliar en la imagen.

1.7 Limitaciones

- No se identifican patologías, desordenes fisiológicos, ni carencia de nutriente presentes en las hojas del jitomate.
- El modelo localiza las hojas y no elimina la presencia de insectos, impurezas (tierra, telarañas, necrosis, etc.) áreas enfermas u objetos dentro de la lámina foliar segmentada.

- No se cuantifica el área de la hoja localizada.

1.8 Justificación

Para el año 2030 se estima que la demanda mundial aumentará de 8.92 a 11.78 millones de toneladas métricas (*MMt*) lo que supone un crecimiento acumulado de 32.10%, mientras que la producción nacional de jitomate estima una expansión de 3.35 a 7.56 *MMt*, lo que representa un crecimiento acumulado de 125.80% (SAGARPA, 2017).

Ante estas tendencias manifestadas por parte de la comunidad agrícola y las tasas con tendencia al alza, surge la oportunidad de desarrollar e implementar un SVA en el área, con el fin de ayudar a los campesinos en las tareas como el monitoreo de las plantas del jitomate, que es una tarea ardua, pero esencial para preservar la salud y la calidad del cultivo y así garantizar las demandas futuras.

El desarrollo de esta aplicación para la segmentación de imágenes capturadas en entornos naturales complejos contribuirá a que otros SVA implementados en la agricultura de precisión operen de manera eficiente, ya que la etapa de segmentación es una etapa clave para delimitar con exactitud el área de interés cada ejemplar y eliminar la presencia de elementos de fondo no relevantes; la correcta ejecución ayuda a reducir errores de inferencia en etapas posteriores del *pipeline* de monitoreo, durante su operación en campo abierto.

1.9 Hipótesis

La modificación del *backbone* de la red *Yolov11n* mediante la integración de los módulos *MV2Block* y *MobileViTBlock*, en combinación con el reemplazo de la función de pérdida por regresión *CloU* por la función *WIoUv3*, permitiría obtener un modelo (*YoloLeaf*) que incrementa el *Recall* en la localización de hojas de jitomate bajo condiciones adversas de campo abierto y reduce la complejidad computacional en comparación con las redes *Yolov11n* y *Yolov12n* estándar.

1.10 Método de solución

En la Figura 1.3 se esquematiza el flujo de trabajo metodológico (*pipeline*) diseñado para localizar y segmentar las hojas de las plantas de jitomate, analizando imágenes adquiridas en condiciones naturales de campo no controladas o adversas.

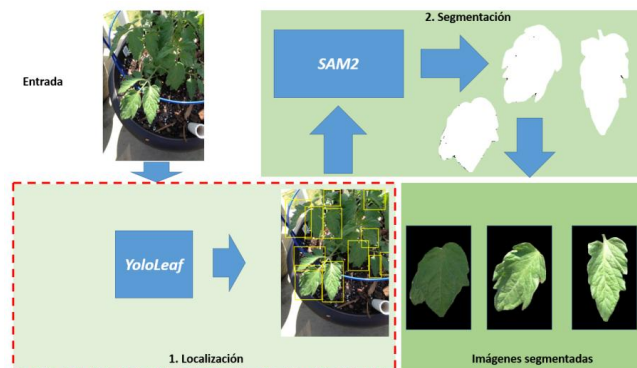


Figura 1.3. Etapas del método propuesto.

La principal contribución de esta investigación reside en la fase de “localización” al proponer una configuración estructural de la red *Yolov11n* (Rahima Khanam & Muhammad Hussain, 2024). Para ello se realizaron modificaciones en la arquitectura y un remplazo de la función original *CIoU* con el fin de minimizar errores de inferencia al localizar hojas que presentan variaciones multiescala y morfología debido a los factores de estrés o condiciones fitosanitarias previamente mencionados.

Para mejorar la extracción de características de las hojas y diferenciarlas de otros elementos del cultivo como tallo, frutos no maduros, flores y otros objetos verdes, se propuso modificar la topología de extracción de características (*backbone*) al remover las capas siete, ocho y diez basados en operaciones convolucionales por los módulos propuestos *MV2Block* y *MobileViTBlock*. *MV2Block* genera mapas de características a un bajo costo computacional (Hüseyin et al., 2022) (Sachin & Mohammad, 2022) y, al mismo tiempo se hace una selección estratégica de las características en cada canal mediante *CBAM* (Woo et al., 2018) que enfatiza y suprime información en los mapas. *MobileViTBlock* calcula relaciones espaciales globales empleando *ViTs* y, la fusiona con información local para definir un mapa refinado que mejora la representación de los datos.

El propósito del remplazo de la función original *CIoU* se considera porque en (Tao J et al., 2025) se intercambia la función original *CIoU* de la red *Yolov11n* por *Wise-IoU loss function* para una mejor predicción en casos donde las oclusiones ocurren a diferentes escalas. La versión *WIoUv3* cuenta con un mecanismo de atención dinámico no monótono que evalúa de forma estratégica la calidad de la caja de anclaje (*Anchor Box*, *AB*) que define las coordenadas del cuadro delimitador principal (*Bounding Box*) que encasilla las hojas, a partir del mapa de características afinado.

Las modificaciones propuestas en la arquitectura y la sustitución de dicha función minimizan desaciertos de inferencia al considerar aquellas láminas foliares con múltiples escalas y, en entornos donde la iluminación es irregular. Incluso, como se mencionó se reducen problemas de ambigüedades donde figuran otros objetos como flores, jitomates no maduros y hojas anómalas que comparten una tonalidad de color.

El modelo desarrollado “*YoloLeaf*”, acota espacialmente las hojas mediante *Bounding Boxes*. Las coordenadas de las regiones delimitadas se transfieren como entrada primitiva (*prompt*) hacia la arquitectura *SAM2* (Ravi et al., 2024) como guía para realizar una segmentación del área foliar precisa. Este planteamiento permite aprovechar la inferencia genérica de regiones (*zero-shot learning*) inherente al modelo fundacional *SAM2* para segmentar cualquier zona en la imagen (Ravi et al., 2024).

1.11 Contenido

El documento consta de ocho capítulos.

- En el primer capítulo se expone el problema resaltando su complejidad, los objetivos, los alcances y limitaciones, la justificación y, la estrategia propuesta para llegar a la solución.

- El segundo capítulo compila los antecedentes de investigación desarrollados en el acervo del TecNM/CENIDET relacionados con la segmentación de imágenes y, el estudio del estado del arte relacionado con el área de la agricultura de precisión.
- El tercer capítulo desarrolla los fundamentos teóricos relacionados con las técnicas utilizadas en la propuesta de solución.
- El cuarto capítulo detalla la construcción del banco de imágenes de hojas de jitomates y el protocolo de anotación para la creación del *ground truth*.
- El quinto capítulo describe la propuesta metodológica y la modificación del modelo *Yolov11n* diseñado para la localización de hojas bajo condiciones ambientales adversas.
- El sexto capítulo presenta la arquitectura de software, diseño de interfaz y lógica de despliegue de la aplicación móvil resultante.
- El séptimo capítulo analiza las configuraciones indispensables, expone el estudio de ablaciones, el análisis de los resultados obtenidos y, las pruebas finales de localización y segmentación.
- Finalmente, en el octavo capítulo se presentan las conclusiones, delinea el trabajo a futuro y compila las actividades académicas realizadas durante el periodo de estudios.

CAPÍTULO 2

ANTECEDENTES

En este capítulo se compilan algunas tesis desarrolladas en CENIDET (Centro Nacional de Investigación y Desarrollo Tecnológico), relacionadas con los temas de localización, segmentación y aprendizaje profundo. Además, se presenta un análisis del estado del arte vinculado con la agricultura de precisión de los últimos cinco años, centrado principalmente en la detección y/o segmentación de frutos u hojas en imágenes capturadas en exteriores.

2.1 Estudios previos

En el acervo del Tecnológico Nacional de México, campus CENIDET, se encontraron cuatro trabajos enfocados en la resolución de problemas relacionados con la temática de segmentación de imágenes durante el segundo semestre del año 2022:

- “Sintonización de una Red Totalmente Conectada para Segmentación de Dos Clases de Objetos en Imágenes” presentado por Diego Gabriel Suarez Santiago (Suarez, 2018).
- “Segmentación no Paramétrica de Tejidos Cerebrales Mediante una Arquitectura Paralela de Redes Neuronales Convolucionales” presentado por Mildred Morales Xicohtencatl (Morales, 2018).
- “Sobresegmentación de Imágenes a Color Mediante un Grafo Difuso Basado en una Medida de Distancia Espacial y Color” presentado por Diego Carlos Luna Márquez (Luna, 2021).
- “Sistema de Visión Artificial para la Detección de Plantas Enfermas Mediante Aprendizaje Profundo” presentado por Andros Meraz Hernández (Meraz, 2022).

Las tesis anteriores formulan propuestas metodológicas interesantes para llevar a cabo la segmentación de imágenes en escenarios naturales, a excepción de (Morales, 2018), que trabajó con imágenes médicas mediante arquitecturas de aprendizaje profundo. Sin embargo, en dichas tesis no se aborda el problema de la localización y segmentación de hojas de cultivos, en particular con ejemplares capturados en condiciones naturales adversas y, que empleen redes neuronales convolucionales conjuntamente.

2.2 Estado del arte

El estado del arte de esta investigación se conforma por una recopilación de 69 trabajos. En la Figura 2.1 se presenta una gráfica que ilustra el número de trabajos de las principales editoriales: *Elsevier*, *Frontiers*, *IEEE* y, *MDPI*.

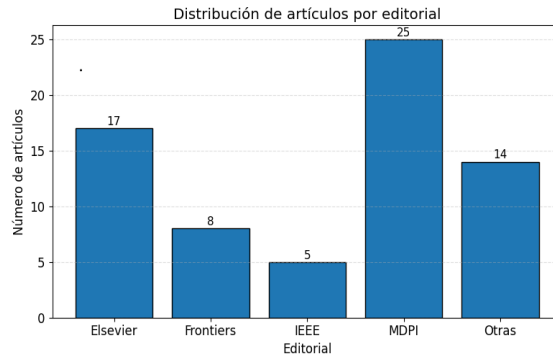


Figura 2.1. Conteo de las principales editoriales.

La editorial *MDPI* concentra el mayor número de trabajos debido a su política de *Open Access* que facilita la descarga del material publicado sin restricción alguna al igual que *Frontiers*. En lo que respecta a *Elsevier* y *IEEE* la disposición de la información es limitada debido al esquema híbrido (se requiere de una membresía).

En la categoría de otras se encuentran publicaciones de congresos como: *Conference: 2023 International Symposium on Networks, Computers and Communications (ISNCC)* y, *Pacific-Rim Symposium on Image and Video Technology (PSIVT)*.

Además de otras revistas como: *International Journal of Advanced Computer Science and Applications (IJACSA)*, *Scientific Reports (Nature)* y, *Engineering, Technology & Applied Science Research (ETASR)*.

En la Tabla 2.1 (pp. 9 - 14), se presenta un condensado de investigaciones de los últimos cinco años que abordan los problemas de localización, principalmente de frutas y hojas bajo múltiples escenarios que se suscitan a campo abierto.

El criterio de selección se basa principalmente en las estrategias de los trabajos que contemplan redes neuronales profundas, especializadas en tareas de localización y/o segmentación del objeto bajo estudio. Adicionalmente, se priorizan aquellas publicaciones en donde se efectúan modificaciones en la topología de la red y/o remplazos de alguna función de regresión que demuestren un beneficio cuantitativo en el desempeño final del modelo.

Tabla 2.1. Síntesis del estado del arte relacionado a la localización y/o segmentación de cultivos.

Referencia	Objetivo	Descripción	Dataset	Distribución de imágenes	Resultados
Afzaal et al., 2021	Localizar y segmentar hojas anómalas de las plantas de fresas.	<i>Mask R-CNN</i> (modificación). Se propone remplazar el <i>backbone</i> original, utilizando <i>ResNet50</i> y <i>ResNet101</i>. El mejor resultado para localizar se registra con <i>Mask R-CNN</i> y <i>ResNet50</i>.	Se adquieren 2,500 ejemplares provenientes de invernaderos y campos de cultivo.	58%, 12.30%, 29.70% Entrenamiento Validación Prueba	72.06% <i>mAP</i> .

Tabla 2.1. Síntesis del estado del arte relacionado a la localización y/o segmentación de cultivos.

Referencia	Objetivo	Descripción	Dataset	Distribución de imágenes	Resultados
Huang et al., 2021	Localizar y segmentar daños en la hoja del maíz.	<i>Yolact++</i> (modificación). Se propone integrar el módulo <i>CBAM</i> a la red <i>Yolact++</i> para enfatizar las características más relevantes y mejorar el rendimiento del modelo, centrándose en las áreas enfermas de las hojas de maíz.	Se recolectan 1,200 ejemplares con un UAV (<i>unmanned aerial vehicle</i>) directamente de los sembradíos.	50%,25%,25% Entrenamiento Validación Prueba	98.71% <i>Precision</i>
Tassis et al., 2021	Identificar enfermedades y plagas en las hojas de la planta del café.	Se utiliza la red <i>Mask R-CNN</i> para localizar y segmentar las láminas foliares.	<i>BRACOT</i> aloja un total de 300 imágenes de hojas de café Arábica en condiciones naturales adversas.	No especifica	73.90% <i>Precision</i> , 71.90% <i>Recall</i>
Hamish A. et al., 2022	Identificar anomalías en las hojas del maíz.	Se emplea <i>Mask R-CNN</i> para localizar y segmentar el área foliar.	Se obtienen 2,332 muestras digitales a partir de visitas a los campos y <i>PlantVillage</i> .	75%,25% Entrenamiento Validación	99.0% <i>Precision</i> y <i>Recall</i>
Storey et al., 2022	Detectar patologías en hojas del árbol de manzanas.	<i>Mask R-CNN</i> (modificación). Se cambia el <i>backbone</i> original con las redes: <i>ResNet-50</i> , <i>MobileNetV3-Large</i> y <i>MobileNetV3-Large</i> .	Se reúnen 3,642 imágenes provenientes de <i>dataset (Plant Pathology, 2020)</i> .	50%,50% Entrenamiento Validación	48.60% <i>AP @IoU50</i> , 28.20% <i>AP @IoU75</i>
Wu et al., 2022	Detectar los frutos de <i>Camellia Oleífera</i> .	Estudio comparativo del desempeño de cinco modelos (cuatro de la familia <i>Yolo</i> y <i>Faster R-CNN</i>). Los mejores resultados se logran utilizando la técnica <i>Mosaic data augmentation</i> con la red <i>Yolov7</i> .	Se recolectan 837 imágenes en los campos de cultivo. El tamaño del banco de imágenes con aumento de datos tradicionales (rotaciones, efecto espejo horizontal y vertical, aumento y reducción de iluminación y ruido gaussiano) es de: 5,854. El conjunto total de imágenes con aumento de datos	90% 10% Entrenamiento Validación	94.76% <i>Precision</i> , 95.54% <i>Recall</i> , 95.15% <i>F1</i> , 96.03% <i>mAP</i>

Tabla 2.1. Síntesis del estado del arte relacionado a la localización y/o segmentación de cultivos.

Referencia	Objetivo	Descripción	Dataset	Distribución de imágenes	Resultados
			elaborado con la técnica <i>Mosaic data augmentation</i> es de: 5,248.		
Yao <i>et al.</i> , 2022	Clasificar dos tipos de enfermedades que ocurren en las hojas de la planta del kiwi.	Se entrena YOLOX para efectuar la localización de los objetivos.	1,000 imágenes adquiridas bajo las condiciones naturales del exterior.	70%,30% Entrenamiento Validación	95% <i>mAP</i>
Xudong <i>et al.</i> , 2022	Detectar dos tipos de patologías en hojas de la planta de papa.	La red <i>Mask R-CNN</i> se utiliza para la localización de hojas en ambientes naturales complejos. Además, se sustituye el <i>backbone</i> de <i>Mask R-CNN</i> por <i>ResNet50</i> y <i>ResNet101</i> . El mejor desempeño se logra con <i>Mask R-CNN</i> y <i>ResNet101</i> .	Se recolectan 500 ejemplares. La distancia del objetivo con respecto al lente de la cámara es de 50 cm.	80%,20% Entrenamiento Validación	97.13% <i>Precision</i> , 81.87% <i>AP</i>
Khalid <i>et al.</i> , 2023	Detectar tres tipos de plagas pequeñas en los cultivos.	Evaluación comparativa de cinco modelos contra <i>Yolov8n</i> .	9,875 capturas digitales tras el aumento de datos (sin especificar).	80%,10%,10% Entrenamiento Validación Prueba	84.70% <i>mAP</i>
Khan F <i>et al.</i> , 2023	Detectar y clasificar enfermedades en los cultivos de maíz.	Análisis del desempeño de cinco modelos de la familia <i>Yolo</i> con diferentes configuraciones de hiperparámetros en cada red. <i>Yolov8n</i> logra los mejores resultados.	Se recolectan 2,675 imágenes bajo condiciones naturales de campo. La distancia del objetivo y de la cámara es de 33 cm.	80%,20% Entrenamiento Validación	99.04% <i>mAP</i>
Lakhdari <i>et al.</i> , 2023	Detectar enfermedades en diferentes cultivos.	Evaluación del desempeño de las redes <i>Faster R-CNN v2</i> , <i>Faster R-CNN v3</i> y <i>Yolov8l</i> .	No se especifica claramente el número total de ejemplares. 68.30% corresponde a hojas de cultivos diferentes. 14.50% y 17.20% son imágenes de jitomate sano y enfermo respectivamente.	91.60%, 4.70% 3.70% Entrenamiento Validación Prueba	60% <i>mAP50</i>
Liu, W <i>et al.</i> , 2023	Detectar nueve patologías en hojas de la	<i>Yolox</i> (modificación) Se reemplaza el <i>backbone</i> original por	Se utilizan 18,834 ejemplares de <i>PlantVillage</i> .	80%,10%,10% Entrenamiento Validación	98.56% <i>mAP</i>

Tabla 2.1. Síntesis del estado del arte relacionado a la localización y/o segmentación de cultivos.

Referencia	Objetivo	Descripción	Dataset	Distribución de imágenes	Resultados
	planta del jitomate.	<i>MobileNetV3</i> como un extractor de características más eficiente. se adopta la función de pérdida de entropía cruzada adaptativa para mitigar efectos negativos del desbalanceo de imágenes.	Se emplea <i>CycleGAN</i> para generar nuevos ejemplares y corrección de calidad, logrando un aumento de 67,457 capturas.	Prueba	
Shinoda <i>et al.</i> , 2023	Localizar y segmentar el fruto de jitomate.	<i>MASK R-CNN</i> (modificación). Remplazar el <i>backbone</i> original con: <i>ResNet50</i> , <i>ResNet101</i> , <i>Swin-T</i> y <i>Swin-S</i> . El mejor resultado de localización se logra con <i>Mask R-CNN</i> + <i>Swin-S</i> .	642 imágenes capturadas directamente en los campos de cultivo. Las transformaciones realizadas son: desenfoques y ruido Gaussiano. No se especifica el número de ejemplares creados.	83%, 17% Entrenamiento Validación	70.60% AP 77.70% AR
Slimani <i>et al.</i> , 2023	Detectar enfermedades en los cultivos de habas.	Evaluación comparativa de seis modelos de la familia <i>Yolo</i> contra <i>Yolov8l</i>	El <i>dataset</i> aloja 3,296 Imágenes después de aplicar: rotaciones, acercamientos, iluminación y saturación.	No especifica	95.10% Precision, 89.50% Recall, 93.70% mAP, 76.50% mAP@50-95
Xia <i>et al.</i> , 2023	Detectar diferentes tamaños del fruto Kiwi.	<i>Yolov8</i> (modificación) Se realizan los cambios en el <i>backbone</i> y <i>neck</i> de la red, implementando <i>Swin Transformer</i> .	Se obtienen 3,000 muestras digitales en entornos naturales complejos. Tras el aumento de datos se generan 5,287 imágenes, aplicando sólo efecto espejo tanto en horizontal y vertical.	70%,20%10% Entrenamiento Validación Prueba	95.10% Precision, 93.20% Recall, 94.70% mAP@50, 71.20% mAP@50-95
Abdullah <i>et al.</i> , 2024	Detección de nueve patologías en hojas de jitomate.	Evaluación del desempeño de la red <i>Yolov8s</i> .	<i>PlantVillage</i>	80%,10%,10% Entrenamiento Validación Prueba	93.20% Precision, 90.30% de Recall 92.50% de mAP,
Wang Y <i>et al.</i> , 2024	Detección de seis afecciones en	<i>Yolov6</i> (modificación) Se integra el módulo de atención <i>CBAM</i>	<i>PlantVillage</i> , <i>PlantDoc</i> y <i>Tomato biotic stress</i>	80%, 20% Entrenamiento Validación	92.90% Precision, 95.20% Recall, 94% F1 93.80% mAP

Tabla 2.1. Síntesis del estado del arte relacionado a la localización y/o segmentación de cultivos.

Referencia	Objetivo	Descripción	Dataset	Distribución de imágenes	Resultados
	hojas de jitomate.		<i>classification dataset</i>		
Xiao <i>et al.</i> , 2024	Detección de tres patologías en hojas de jitomate.	<i>Yolov8</i> (modificación) Se integran: <i>GIP-SPPF</i> para optimizar la información global de los bordes y así reducir el desvanecimiento de la gradiente. <i>KOIDSConv</i> para mejorar la fusión de información a un menor costo computacional.	445 imágenes originales. 4,895 aplicando aumento de datos	No especifica	98.06% <i>mAP@50</i>
Abulizi <i>et al.</i> , 2025	Detección de ocho enfermedades en hojas de jitomate.	<i>Yolov9</i> (modificación) Se incorpora el módulo <i>DySample</i> para optimizar la extracción de características a diferentes escalas.	4,124 imágenes procedentes de entornos controlados y no controlados.	80%,10%,10% Entrenamiento Validación Prueba	92.50% <i>Precision</i> , 86.60% <i>Recall</i> , 95.10% <i>AP</i> , 86.40% <i>mAP</i>
Kaur <i>et al.</i> , 2025	Detectan seis patologías en hojas de los cultivos: jitomate, papa, maíz y uva.	<i>Yolov8</i> (modificación) <i>Batch Normalization</i> estabiliza y acelera el proceso de entrenamiento, conllevando a una mejora en la convergencia y desempeño. <i>Dropout Layer</i> evita problemas de sobreajuste ya que, regulariza estratégicamente la activación de las neuronas durante el entrenamiento.	<i>PlantDoc</i>. <i>Potato disease leaf dataset</i>. <i>Corn or Maize leaf disease dataset</i>. <i>PlantVillage</i> <i>Grapevine disease images</i>. se crean 23,032 ejemplares tras aplicar: rotaciones, efecto espejo, acercamiento, ajuste en la iluminación y recortes.	80%,10%,10% Entrenamiento Validación Prueba	98.50% <i>Precision</i> , 98.00% <i>Recall</i> , 99.90% <i>mAP@50</i> 94.00% <i>mAP@50-95</i> .
Liu,Z <i>et al.</i> , 2025,	Detección de ocho afecciones en hojas de la planta del jitomate.	<i>Yolov8s</i> (modificación) El bloque <i>HCMamba</i> mejora la representación de características. La capa <i>BiFPN</i> se emplea para reemplazar el componente de fusión de características original, contribuyendo a una	<i>PlantVillage</i>, <i>PlantDoc</i> y <i>Taiwan Tomato Leaf Disease</i> Se reúnen 3,362 ejemplares. Se crean 2,354 imágenes tras aplicar: rotaciones, ajustes de brillo, correcciones en escala de grises	70%,20%,10% Entrenamiento Validación Prueba	81.90% <i>F1</i> 86.70% <i>mAP@0.5</i> 72.00% de <i>mAP@0.5:0.95</i>

Tabla 2.1. Síntesis del estado del arte relacionado a la localización y/o segmentación de cultivos.

Referencia	Objetivo	Descripción	Dataset	Distribución de imágenes	Resultados
		integración de características eficiente a diferentes escalas.	y se añade ruido gaussiano, sólo en la partición de entrenamiento.		
Tao J <i>et al.</i> , 2025	Detectan cuatro enfermedades en hojas del árbol de manzanas.	<i>Yolov11n</i> (modificación) El bloque <i>CWSCConv</i> reduce el costo computacional, mejorar la detección de objetos pequeños é inferencia del modelo ante la amplia diversidad de cambios físicos de la hoja y fondos complejos. El nuevo módulo <i>C2PSA_ECCAttention</i> realiza una captura más precisa de las características de enfermedades en las hojas. La capa <i>FMLAttention</i> realiza la obtención de características finas y detalles locales relevantes. Aplica <i>Wise-IoU loss function</i> para una mejor predicción en casos de oclusiones a diferentes escalas.	2,452 imágenes capturadas directamente del campo. 9,808 muestras aplicando: rotaciones, traslaciones, ajustes en el brillo y efectos espejos.	80%,10%,10% Entrenamiento Validación Prueba	85.50% <i>Precision</i>, 81.20% <i>Recall</i>, 87.30% <i>mAP@50</i> 57.10% <i>mAP@50-95</i>
Wang X y Liu J, 2025	Detectar cuatro enfermedades en el jitomate en hojas y frutos.	<i>Yolov10</i> (modificación) El módulo <i>MPIRU</i> mejora la extracción y fusión de características en diferentes escalas. El bloque <i>DFAF</i> se enfoca adaptativamente en las regiones clave de las enfermedades. La función <i>Focal-EIoU</i> que mitiga el problema de clases desbalanceadas y favorece a una estimación de los <i>Bounding Boxes</i> precisa.	10,537 ejemplares capturados en ambientes naturales complejos. Se crean 31,053 muestras con: ajustes de iluminación, rotaciones, efecto espejo en horizontal y vertical.	80%,10%,10% Entrenamiento Validación Prueba	94.23% <i>mAP@50</i> 72.52% <i>mAP@50-95</i>

2.3 Conclusiones del estado del arte

En algunos artículos consultados se opta por simplificar el entorno externo de captura donde el punto focal suele estar dirigido hacia la(s) hoja(s), evadiendo la digitalización de otros elementos como en (Xudong *et al.*, 2022) donde la distancia del objetivo y la lente es de 50 cm y en (Khan F *et al.*, 2023) 33 cm, omitiendo la complejidad multiescala inherente a escenarios agrícolas no estructurados.

La familia de redes neuronales *Yolo* es la más utilizada para efectuar detecciones de enfermedades (localización y clasificación) donde figuren frutos o vegetales (Wu *et al.*, 2022), (Xia *et al.*, 2023) y, (Wang X y Liu J, 2025) o de manera similar a través de las hojas del cultivo de interés (Abdullah *et al.*, 2024), (Wang Y *et al.*, 2024), (Xiao *et al.*, 2024) y (Abulizi *et al.*, 2025) por mencionar algunos de los trabajos estudiados.

Al igual que *Yolo*, las redes *Mask R-CNN* y *Faster Mask R-CNN* también son implementadas para tareas de detección de patologías en hojas (Afzaal *et al.*, 2021), (Tassis *et al.*, 2021) o frutos (Lakhdari *et al.*, 2023) y. (Shinoda *et al.*, 2023). Sin embargo, las variantes de *Mask R-CNN* y *Faster Mask R-CNN* requieren un mayor poder computacional comparado con las variantes *Yolov11n* (Tao J *et al.*, 2025) y *Yolov8s* (Liu,Z *et al.*, 2025) que son adecuadas para su ejecución en equipos con pocos recursos de *hardware*.

Debido a la practicidad de la red *Yolo*, en la literatura se encuentra varias publicaciones donde se utilizan versiones mejoradas para un fin. Por ejemplo, en (Yao *et al.*, 2022) se presenta un método que combina tres redes para la localización, segmentación y clasificación. *Yolox* (versión mejorada de *Yolov3*) se entrena para estimar la ubicación de hojas de la planta del kiwi; *DeepLapWebv3* para segmentar áreas anómalas en la hoja previamente localizada; *ResNet101* para clasificar la enfermedad a partir de las lesiones segmentadas en la etapa anterior. No obstante, el trabajo de (Huang *et al.*, 2021) se presenta una versión mejorada de la red *Yolact++* que se especializa en tareas de localización y segmentación, una alternativa que simplifica las etapas de un esquema de trabajo.

En (Huang *et al.*, 2021) se modifica la red *Yolact++* al integrar el módulo de atención *CBAM* en la red de extracción de características (*backbone*) para mejorar la detección de lesiones pequeñas y difusas. Además, se optimiza la red de pirámide de características (*FPN*) para fusionar mejor las capas de información, logrando una precisión de segmentación del 98.71% en hojas de maíz.

Otros trabajos que se basan en remplazar el *backbone* original por otras estructuras de redes convolucionales como en (Afzaal *et al.*, 2021) donde se propone remplazar el extractor de características original, utilizando *ResNet50* y *ResNet101*, para mejorar la localización y segmentación de hojas enfermas de la planta de fresa. El mejor resultado que define la *ROI* es de 72.06% *mAP* con la configuración *Mask R-CNN* + *ResNet50*. En (Storey *et al.*, 2022), se efectúa el procedimiento de eliminar el *backbone original* y sustituirlo con *ResNet-50* o *MobileNetV3-Large* para la detección de patologías en hojas del árbol de manzanas. La fusión *Mask R-CNN* + *ResNet-50* es seleccionada como la mejor para segmentar y detectar enfermedades en las láminas foliares logrando 48.60% *AP @IoU50* y 28.20% *AP @IoU75*.

En distintas investigaciones, la sustitución del *backbone* también se efectúa al usar transformadores de visión como en (Shinoda *et al.*, 2023) donde se hace un estudio comparativo de cambio del extractor de características con: *ResNet50*, *ResNet101* vs *Swin-T* y *Swin-S* de la red base *MASK R-CNN* para la localización y segmentación del jitomate (fruto). El mejor resultado de localización del objetivo se logra con *Mask R-CNN + Swin-S* al puntuar 70.60% de *AP*. De manera semejante en (Xia *et al.*, 2023) su método consiste en modificar tanto el *backbone* como el *neck* de la red *Yolov8* al implementar módulos de atención basados en *Swin Transformer* para la detección precisa de los diferentes tamaños del fruto kiwi. Los mejores resultados que se reportan son 95.10% *Precision*, 93.20% *Recall*, 94.70% *mAP@50*, y 71.20% *mAP@50-95*.

Además de realizar cambios en la red principal, en (Tao J *et al.*, 2025) y (Wang X & Liu J, 2025), las modificaciones son avanzadas ya que, se remplace la función de regresión original que estima la localización de los objetivos. (Tao J *et al.*, 2025) intercambia la función original *CIoU* de la red *Yolov11n* por *Wise-IoU loss function* para una mejor predicción en casos de oclusiones a diferentes escalas. (Wang X & Liu J, 2025) retira la función original de *Yolov10* e implementa la función *Focal-EIoU* que mitiga el problema de clases desbalanceadas y favorece a una estimación de los *Bounding Boxes* precisa.

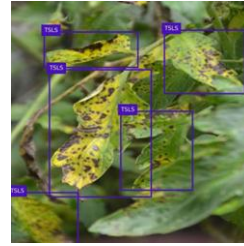
Para la creación de un modelo predictivo eficiente, se requiere de una correcta selección de datos (en este caso imágenes) que contribuya a una clara descripción del problema a resolver. Por ejemplo, en los artículos se centran en escenarios controlados como en (Abdullah *et al.*, 2024), que utilizan ejemplares digitales provenientes de *PlantVillage*, donde figuran únicamente una hoja del jitomate colocada sobre una superficie que contrasta la lámina foliar. Los resultados que se reporta son: 93.20% de *Precision*, 90.30% de *Recall* y 92.50% de *mAP*, detectando 10 clases (una clase sana y nueve patologías). En contraste, (Abulizi *et al.*, 2025) tiene como objetivo localizar y clasificar nueve estados de salud de hojas de la planta del jitomate (ocho enfermedades y la categoría saludable) con ejemplares capturados en entornos no controlados. Los mejores resultados reportados son: 92.50% *Precision*, 86.60% *Recall*, 95.10% *AP* y 86.40% *mAP*.

Otras investigaciones como (Wu *et al.*, 2022), (Xia *et al.*, 2023), (Wang *et al.*, 2024) y (Liu, Z *et al.*, 2025) combinan imágenes de distintas fuentes como laboratorios, invernaderos y a campo abierto con el fin de crear un banco de imágenes que represente las innumerables circunstancias que ocurren en diferentes situaciones. Adicionalmente, aplican el proceso de aumento de datos tradicional (con operaciones geométricas y fotométricos) para enriquecer el contenido y lograr una mejor generalización del modelo. Por su parte (Liu, W *et al.*, 2023) aplica una técnica avanzada de aumento de datos al generar datos sintéticos utilizando redes generativas antagónicas.

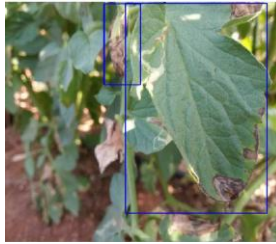
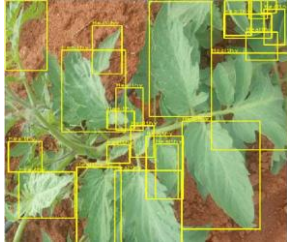
En la Figura 2.1, se ejemplifican las características de las hojas que constituyen los bancos de imágenes de algunos trabajos.



Ejemplos de (Abulizi et al., 2025)



Ejemplos de (Wang et al., 2024)



Ejemplos de (Xiao et al., 2024)

Observe que el punto focal es dirigido a la(s) láminas foliare(s).

Figura 2.2. Evidencias de los bancos de imágenes.

CAPÍTULO 3

MARCO TEÓRICO

En este capítulo se presentan los fundamentos teóricos en el dominio agrícola y computacional vinculados al desarrollo de la investigación. Inicialmente se describe el cultivo de interés y algunas patologías que alteran las características físicas de las hojas de la planta. Posteriormente, se aborda la información de la red neuronal convolucional *Yolo* y los mecanismos de atención basados en Transformadores de Visión, las funciones de regresión orientadas a tareas de detección y las métricas de evaluación para medir el desempeño del modelo.

3.1 Cultivo bajo estudio

a) Jitomate

El jitomate o tomate (*Lycopersicon esculentum* Mill) es una planta de la familia de las solanáceas que produce las regiones de América del Sur. Necesita de una luz intensa (70,000 luxes como punto de saturación), gusta de un clima fresco con gran diferencia de temperaturas diurnas y nocturnas, y no es resistente a la humedad excesiva. Desde la época prehispánica es cultivada y, actualmente en las zonas áridas de México es el cultivo más comúnmente desarrollado (López et. al, 2004).

El tomate es una planta anual que en condiciones favorables vive y fructifica varios años. No obstante, se prefiere la siembra anual por razones de carácter práctico. A continuación, se enlistan y se describen algunas de sus características botánicas, de acuerdo con (López et. al, 2004):

Sistema de raíces: Algunas raíces alcanzan hasta 150 cm de longitud. La mayor parte de las raíces se desarrollan en la capa del suelo que se extiende desde 5 cm hasta 60-70 cm de profundidad.

Tallo: El tallo de las plantas jóvenes de tomate es cilíndrico, más tarde se hace angular. Está cubierto de finos vellos; unos más largos, otros más cortos. Los últimos segregan una sustancia viscosa de color verde oscuro y olor específico para el tomate.

Se tiene conocimiento de dos variedades en el desarrollo de la planta: indeterminada y determinada (Figura 3.1). En el primer caso el tallo principal y todas sus ramificaciones terminan en racimo floral; el crecimiento vertical de las plantas es limitado. El segundo caso, el racimo floral ya terminado en el seno de la última hoja forma un hijo que prosigue el crecimiento del tallo principal. Por lo tanto, el crecimiento vertical no está limitado por las características morfológicas de la planta.

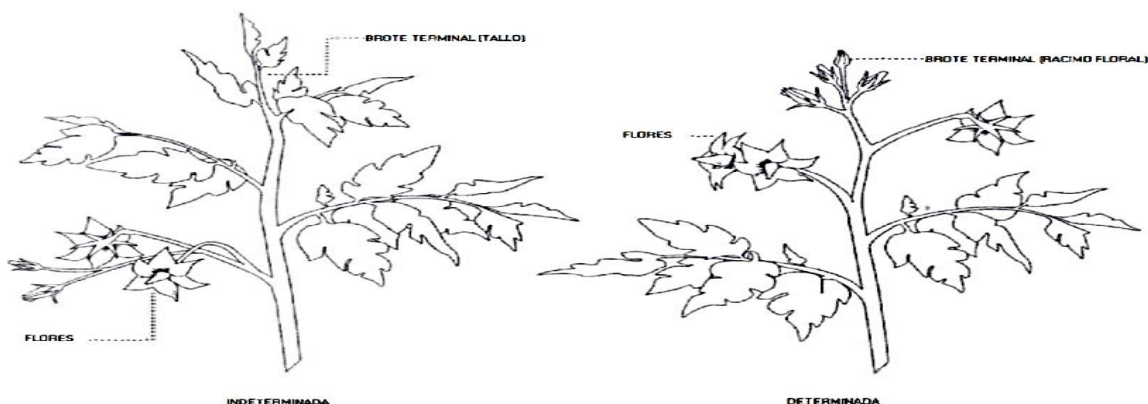


Figura 3.1. Tipos de tomate según su hábito de crecimiento (López. *et. al*, 2004).

Según las condiciones ambientales y el modo de cultivo, el tallo alcanza una altura de hasta 4 m.

Hojas: Son alternas, varían mucho en cuanto a tamaño, las peculiaridades del margen y carácter de la superficie, según las características genéticas de las variedades y condiciones del cultivo.

Presenta de siete a nueve folíolos peciolados que miden 4-60 mm x 3-40 mm, lobulados y con borde dentado, alternos, opuestos y, por lo general, de color verde, glanduloso-pubescente por el haz y ceniciento por el envés. Se encuentra recubierta de pelos glandulares y dispuestos en posición alternada sobre el tallo (Monardes 2009). La posición de las hojas (enana, hoja de papa, estándar, *peruvianum*, *pimpinellifolium* o *hirsutum*) en el tallo puede ser semierecta, horizontal o inclinada (IPGRI 1996).

En la Figura 3.2 se ilustran ocho variaciones naturales de la planta del jitomate. De izquierda a derecha se presentan las hojas junto con su inflorescencia respectiva: *S. peruvianum* (LA0153), *S. neorickii* (LA0247), *S. pimpinellifolium* (LA0373), *S. lycopersicum* var. *cerasiforme* (LA0292), *S. chilense* (LA1930), *S. chmielewskii* (LA1028), *S. pennellii* (LA706), *S. lycopersicum* cv. M82 (LA3475) (Zsögön *et. at.*, 2017).



Figura 3.2. Diferentes especies naturales del jitomate (Zsögön *et. at.*, 2017).

b) Agentes patógenos

Los agentes patógenos o fitopatógenos son organismos que causan enfermedades en las plantas que alteran su fisiología normal, afectando su crecimiento, su reproducción y, en el peor de los casos, causar su muerte (Agrios, 2005). La ciencia los clasifica en cuatro grandes grupos principales (Lucas, 2020; Schumann & D'Arcy, 2012):

- **Bacterias**

Son organismos unicelulares que suelen entrar a la planta a través de heridas (causadas por poda, granizo o insectos) o por aberturas naturales como las estomas (Lucas, 2020).

- **Hongos**

Son responsables de aproximadamente el 80% de las enfermedades en las plantas. Se propagan fácilmente a través de esporas transportadas por el viento, el agua o insectos (Alexopoulos et al., 1996).

- **Nematodos**

Son pequeños gusanos microscópicos que habitan principalmente en el suelo y atacan el sistema radicular (Perry & Moens, 2013).

- **Virus**

Son partículas microscópicas que no se multiplican por sí solas; necesitan de las células de la planta para replicarse. Generalmente son transmitidos por insectos vectores, como pulgones o mosca blanca (Matthews, 2012).

A continuación, se describen tres patologías causadas por dichos microorganismos en la planta del jitomate, resaltando principalmente los síntomas en las hojas.

Bacterial spot (mancha bacteriana): Es una enfermedad causada por la bacteria: *Xanthomonas spp* en el cultivo de tomate. Esta enfermedad es difícil de controlar cuando las condiciones climáticas apoyan a su multiplicación. Las precipitaciones abundantes y las temperaturas cálidas favorecen el desarrollo de la enfermedad. El agente patógeno se prolifera en semillas y/o por una mala manipulación cuando el cultivo ya presenta la enfermedad (MSU, s.f.). Los síntomas que identifican a este padecimiento en el cultivo son: las hojas con muchas lesiones desarrollan una apariencia amarillenta. A medida que crecen las lesiones pueden causar daño o una destrucción rápida y extensa. El follaje muerto puede permanecer en la planta. La Figura 3.3 presenta una fotografía con la enfermedad mancha bacteriana.



Figura 3.3. Mancha bacteriana (MSU, s.f.).

Leaf mold (hoja con moho): Es un padecimiento común en los cultivos de tomates causado por los hongos: *Fulvia fulva*, formerly *Cladosporium fulvum* y *Passalora fulva*. Los niveles altos de humedad y temperaturas templadas contribuyen al desarrollo de la enfermedad. El agente patógeno puede sobrevivir en los restos de cultivos por lo que, las esporas son esparcidas por la lluvia, el viento, las herramientas de trabajo, ropa e insectos (MSU, s.f.). Los síntomas que distinguen el estado enfermo del tomate son: desarrollo de

lesiones de color verde pálido o amarillo con bordes irregulares en la superficie superior de los folíolos. Aparición de hongos color verde oliva y textura aterciopelada, que se desarrollan en la parte inferior de los folíolos, directamente debajo de las lesiones amarillas. En la Figura 3.4 se observa la manifestación de la anomalía hoja con moho.



Figura 3.4. Hoja con moho (MSU, s.f.).

Yellow leaf curl virus (Virus de la hoja amarilla): Es una de las enfermedades virales más devastadoras de los tomates, reportando pérdidas totales de cultivos cuando la población del agente es grande. El virus es transmitido por moscas blancas adultas. Las moscas blancas adquieren el virus al alimentarse de una planta infectada en aproximadamente 15 minutos y transmitir el virus después de 6 horas. La transmisión del virus a plantas no infectadas ocurre en aproximadamente 15 minutos de alimentación, ya que las moscas blancas retienen el virus durante varias semanas (MSU, s.f.). Los síntomas más evidentes de esta enfermedad son: provoca un severo retraso en el crecimiento, abscisión de las flores y reducciones significativas en la cosecha. Las hojas son de tamaño reducido y exhiben rizado hacia arriba (ahuecamiento). La Figura 3.5 ilustra una planta de jitomate enferma del virus de la hoja amarilla.



Figura 3.5. Enfermedad del virus de la hoja amarilla (MSU, s.f.).

3.2 Tópicos computacionales

a) Localización

En el ámbito de la visión por computadora, la ubicación de un objeto se define como la tarea algorítmica de identificar la posición y la extensión espacial exacta de un objeto de interés predeterminado dentro de una imagen, bajo el supuesto de que la escena contiene un único elemento principal de dicha categoría (Goodfellow et al., 2016; Szeliski, 2022).

Formalmente, dado un espacio de imágenes I de dimensiones W, H (ancho y alto) con C canales de color, la delimitación espacial se modela como una tarea de regresión multivariada donde una función parametrizada f (generalmente una red neuronal convolucional) mapea la imagen en un espacio continuo de cuatro dimensiones (Russell & Norvig, 2020):

$$f : I \rightarrow \mathbb{R}^4$$

donde la salida es un vector $y = [x, y, w, h]$ que define una caja delimitadora *Bounding Box*):

- (x, y) : Las coordenadas del centro de la caja delimitadora (o, alternativamente, de la esquina superior izquierda).
- w : El ancho (*width*) relativo o absoluto de la caja.
- h : El alto (*height*) relativo o absoluto de la caja.

No se busca clasificar qué objetos están presentes en la escena, se enfoca estrictamente en resolver el componente espacial de un único elemento (Chollet, 2021).

b) Segmentación

En términos matemáticos (ecuación 3.1), la segmentación se define como la fragmentación de una imagen I en N regiones R_i sujeta a una propiedad de homogeneidad (Sonka & Fitzpatrick, 2009).

$$I = \bigcup_{i=1}^N R_i \quad \text{Ecuación 3.1}$$

La homogenización (ecuación 3.2) es una condición esencial que se cumple ya que, esta cualidad permite efectuar la separación de las áreas de una imagen acorde a un criterio basado en características como: color, intensidades, texturas etc.

$$R_i \cap R_j = 0 \mid i \neq j \quad \text{Ecuación 3.2}$$

Es decir, los píxeles que conforman una imagen no comparten fronteras, existe una división correctamente definida.

La Tabla 3.1, presenta tres diferentes técnicas para efectuar la delimitación de áreas en imágenes.

Tabla 3.1. Técnicas de segmentación (Xicohtencatl, 2018).

Píxeles	Fronteras	Regiones
Umbralización	Bordes locales	Fusión y/o división
Histogramas	Bordes globales	Crecimiento de regiones
Agrupamiento	Modelos deformables	Morfológicas
Redes Neuronales		

La segmentación basada en píxeles permite separar un objeto dentro de la imagen del fondo que lo rodea. La técnica se basa en comparar alguna propiedad de la imagen con un umbral fijo o variable, realizando una comparación por cada píxel que conforma a la imagen (Xicohtencatl, 2018).

La segmentación basada en bordes consiste en buscar los cambios bruscos de la imagen que representan una discontinuidad en la característica considerada, por ejemplo, la intensidad de los píxeles (Xicohtencatl, 2018).

La segmentación basada en regiones busca píxeles que tengan características similares, analizando a los píxeles vecinos para verificar si cumplen con el criterio de similitud (Xicohtencatl, 2018).

c) **Deep Learning**

Deep Learning (DL) o aprendizaje profundo (AP) es una subárea del Machine Learning *ML* enfocado en la representación de la información a través de capas. El término “*deep*” hace referencia al número de capas que conforma un modelo de esta naturaleza (Chollet, F, 2018). En la Figura 3.6 se aprecia una arquitectura conformada por cuatro capas.

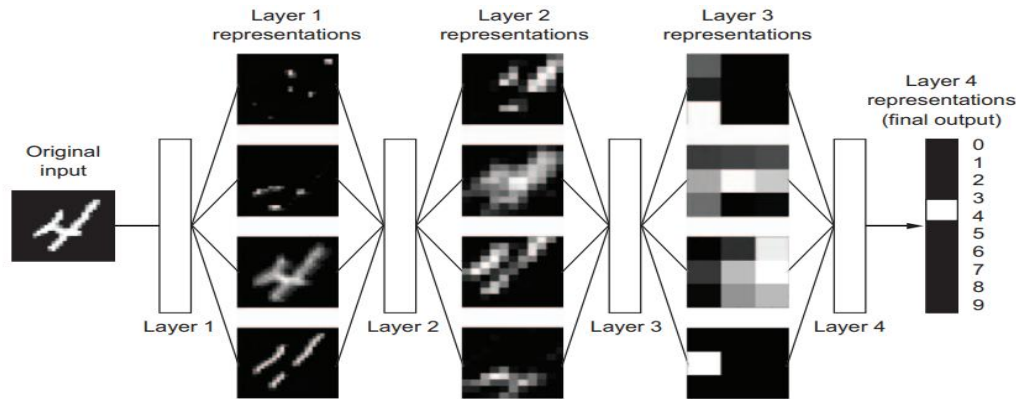


Figura 3.6. Representación del aprendizaje profundo utilizando un modelo de clasificación de imágenes (Chollet, F, 2018).

La red transforma la imagen del número en otra representación totalmente diferente a la fuente original. A medida que los datos pasan a las capas más profundas, la representación es cada vez más notoria. Por lo tanto, el *DL* es técnicamente una manera de aprender la representación de los datos a nivel de multicapas (Chollet, F, 2018).

d) **Ground Truth**

El término *Ground Truth (GT)* describe la información de referencia que se utiliza para: a) el entrenamiento del modelo, donde el contenido es empleado como guía para que la red aprenda cuales características son idóneas para una determinada aplicación y, b) la validación y prueba, donde los registros permiten evaluar cuantitativamente la exactitud del algoritmo entrenado. La adquisición, limpieza y anotación manual de los registros suele ser costosa en términos de tiempo ya que, los modelos de *IA* requieren miles de muestras etiquetadas para generar alcanzar un rendimiento bueno. La información de referencia del *GT* proviene de diferentes modalidades de datos como imágenes, señales o texto (MathWorks, s.f.).

e) **Redes neuronales convolucionales**

La red neuronal convolucional o *Convolutional Neural Network (CNN)* surge como una inspiración directa de los estudios de la corteza visual del cerebro que responsable de procesar la información visual (Tassis *et. al*, 2021). Por ende, las *CNN* son un tipo de arquitectura que se emplea en el campo de reconocimiento de patrones con imágenes (Keiron O'Shea & Ryan Nash, 2015).

Las redes *CNN* surgen debido a que el perceptrón multicapa (un tipo de red neuronal clásico) no es eficiente ya que, al procesar un ejemplar digital cuyas dimensiones son de 300×300 píxeles y, a color (lo que implica tres canales de color para el rojo, azul y verde), si en la primera capa oculta se utilizan únicamente 100 neuronas, el número de parámetros que se tendrán en esta primera operación (sin contar los *bias*) sería de 27 millones. Si se desea utilizar más capas, el número de parámetros crecería rápidamente y por consiguiente el proceso se vuelve ineficiente, explica (Parreño, 2018).

Por el contrario, con las redes *CNN* el costo computacional es menor dado que, a medida que la imagen es procesada a lo largo de la arquitectura, se crean mapas de características y a su vez se reduce la resolución original, conservando sólo aquellos valores relevantes, volviendo el procedimiento eficiente. A continuación, se explican algunos de los operadores más utilizados en las redes *CNN*.

- **Capa convolucional**

Es el núcleo principal de este tipo de arquitecturas de redes neuronales. A partir de una imagen I y mediante la operación convolución se obtienen M mapas de salida que corresponde a cada uno de los filtros utilizados (Parreño, 2018).

La ecuación 3.3 define la función convolucional:

$$Conv_{i,j,1} = \sum_{m=1}^M \sum_{k_1=1}^c \sum_{k_2=1}^c X_{i+k_1-1,j+k_2-1,m} \cdot K_{m(k_1,k_2,m)} \quad \text{Ecuación 3.3}$$

donde:

X : es la imagen de entrada.

K_m : es el filtro convolucional.

$Conv$: es la salida final.

M : número de canales (profundidad).

c : son las dimensiones del filtro.

$i \in [1, W]; j \in [1, H]; m \in [1, M]$

En la Figura 3.7 se ilustra la función convolucional discreta definida en la ecuación 3.3

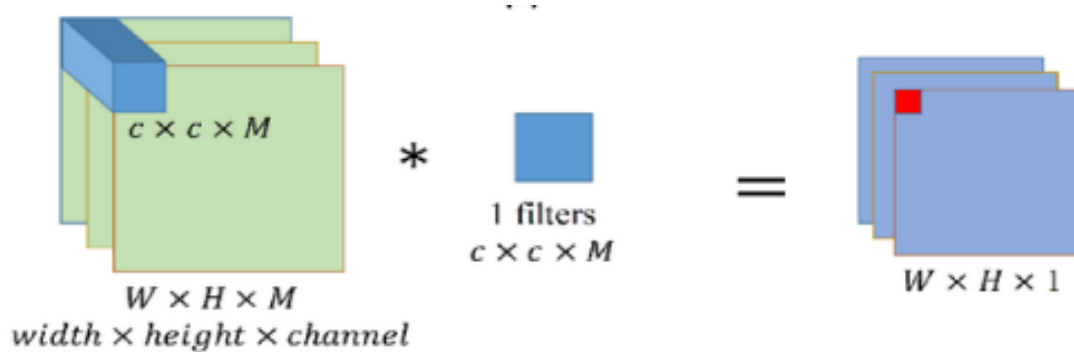


Figura 3.7. Representación gráfica del operador convolucional (Hüseyin *et al.*, 2022).

- **Capa pooling**

Reduce progresivamente el tamaño de los mapas con el fin de disminuir el número de cálculos, y controlar el *overfitting* durante el proceso de entrenamiento. La capa *pooling* trabaja de forma independiente entre los mapas de características ya que, dentro de ella no existen parámetros para aprender, únicamente se calculan medias, máximos, etc., a partir de los mapas de entrada (Parreño, 2018). A continuación, se describen dos tipos de capas de *pooling* más empleados en la práctica:

- **Average Pooling:** El valor de salida es el promedio de los valores de los píxeles dentro del filtro.
- **Max Pooling:** El dato de salida es el máximo de los valores de los píxeles alojados en el *kernel*.

La Figura 3.8 muestra un ejemplo del operador *Max Pooling* cuya ventana (filtro) es de tamaño 2x2 y un *stride* de valor igual a 2.

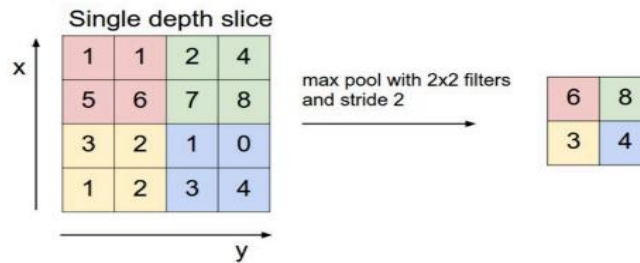


Figura 3.8. Representación gráfica de una operación de *Max Pooling* (Parreño, 2018).

- **Padding y stride**

El *stride* o salto es un cambio de posición horizontal que realiza el *kernel* cada vez que el operador convolucional es aplicado a un mapa de dimensiones $n \times n$, dando como resultado un mapa de menor tamaño, en concreto (Ecuación 3.4):

$$\text{Mapa} = (n - f + 1)(n - f + 1) \quad \text{Ecuación 3.4}$$

En la práctica el salto es un parámetro configurable, usualmente el tamaño que suele emplearse es de un píxel, pero existen modelos que utilizan este salto para disminuir el tamaño de los mapas en vez de utilizar los operadores *poolings*. El tamaño de los mapas de salida finales, incluyendo el *padding* p y el tamaño del *stride* S sería (Ecuación 3.5):

$$\text{Stride} = \left(\frac{(n+2p-f)}{S} + 1 \right) \left(\frac{(n+2p-f)}{S} + 1 \right) \quad \text{Ecuación 3.5}$$

El *padding* o relleno es una técnica que consiste en aumentar el número de filas y columnas en el perímetro de la imagen, asignando el valor de cero a cada casilla. Esta táctica es habitualmente aplicada debido a que, el píxel de la esquina superior izquierda del mapa a procesar sólo se contempla una única vez para el cálculo de los mapas de salida, lo que provoca una posible pérdida de información si se toma en cuenta que un píxel central se contempla hasta en nueve ocasiones. De modo que los mapas de salida quedarían como (expresión 3.6):

$$Mapa = (n + 2p - f + 1)(n + 2p - f + 1)$$

Ecuación 3.6

Lo que daría la posibilidad de tomar en cuenta los valores de los bordes del mapa de entrada, y obtener mapas de salida del mismo tamaño (Parreño, 2018).

- **Capa Reshape**

Esta capa toma como entrada N mapas de características de dimensiones $M \times K$ provenientes de la etapa convolucional de la red y la transforma a un vector de dimensionalidad $N \times M \times K$ con el fin de que, la parte *fully connected* procese la información al tomar las características extraídas por las convoluciones y realizar la tarea de la clasificación (Parreño, 2018).

3.3 Redes neuronales especializadas en detección

a) YOLO (Redmon et al., 2016)

YOLO (You Only Look Once) (versión uno) es una clase de red neuronal convolucional orientada a tareas de detección de objetos (localización y clasificación). Su topología se encuentra conformada por una serie de capas convolucionales (*Conv. layer*), capas de agrupación (*Max pool layer*) subsecuentes (Figura 3.9) ambas, encargadas de generar y seleccionar características de la imagen de entrada. Las capas densas (*Fully connected*) predicen las probabilidades de clase y las coordenadas de ubicación del objetivo.

La arquitectura aloja 24 capas convolucionales, seguida por dos capas densas; las primeras 20 capas convolucionales del modelo se preentrenan con el *dataset* (ImageNet, s.f).

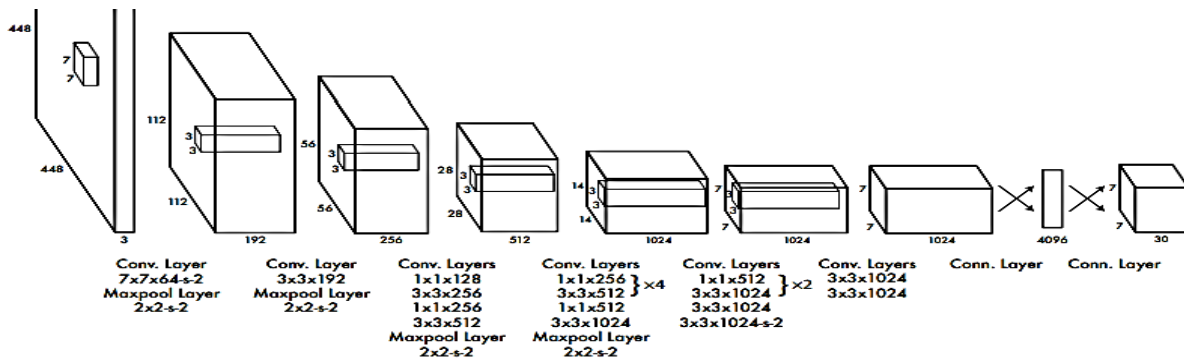


Figura 3.9. Arquitectura YOLO (Redmon et al., 2016).

El proceso interno que ocurre dentro de la red es el siguiente:

- La imagen que ingresa es dividida en varias celdas de dimensiones $S \times S$ (especie de rejilla).
- En cada división, el algoritmo calcula la probabilidad de existencia del objeto y un conjunto de probabilidades de clases $Pr(Class_i | Object)$.
- Un conjunto de celdas define el objeto *BB* (*Bounding Box*) que se integra de cinco predicciones ($x, y, w, h, confidence$); las coordenadas (x, y) representan el centro; el largo (w) y la altura (h); el puntaje de certeza ($Pr(Object) * IoU_{pred}^{truth}$) que se utiliza para determinar la existencia de un objeto dentro de la región.

- d) Si el objetivo no está definido dentro de la celda, el grado de certeza es cero de lo contrario, se toma el valor y se calcula *intersection over union (IOU)* entre la predicción y el *ground truth (GT)*.
- e) Finalmente, el valor calculado del paso anterior, se espera que sea lo más cercano al dato del *GT*.
- f) La predicción final que *YOLO* retorna se compone de: i) las probabilidades de clases definidas y ii) las coordenadas de ubicación del cuadro delimitador principal.

A continuación, se detallan las especificaciones de *Yolov11*, variante de la familia *YOLO* seleccionada como el eje metodológico central de este trabajo de investigación.

b) *Yolov11* (Rahima Khanam & Muhammad Hussain, 2024)

Yolov11 desarrollado por *Ultralytics*, es la tercera versión publicada en 2024. Al igual que sus predecesoras *Yolov11* es un modelo predictivo para efectuar seis múltiples tareas de visión por computadora, cinco de ellas populares como: clasificación, detección, segmentación, seguimiento y estimación de poses del cuerpo humano. *Yolov11* introduce un nuevo enfoque denominado como “detección de objetos orientados”, que permite detectar objetos con respecto a un ángulo determinado, contribuyendo a una mejor precisión de localización objetos rotados. Esta nueva característica es importante en análisis de imágenes aéreas y tareas de automatización en almacenes.

Ya que esta red se basa en *Yolov8*, la arquitectura de *Yolov11* se conforma por tres secciones. El primero es el *backbone* el cual, procesa la información en bruto de la imagen para extraer datos relevantes y así generar mapas de características a diferentes escalas. La siguiente sección llamada *neck* funciona como un proceso intermediario que agrega y mejora las representaciones de las características a través de diferentes escalas. Finalmente, el *head* encargado de realizar las predicciones a partir del mapa de características refinado, generando las cajas delimitadoras y la probabilidad de pertenencia de clases.

Yolov11 incorpora tres bloques nuevos (*C2P2A*, *C3k2* y *CBS*), distribuidos estratégicamente en las tres subdivisiones brevemente descritas. La Figura 3.10, señala los módulos claves y los problemas de visión por computadora que *Yolov11* aborda.

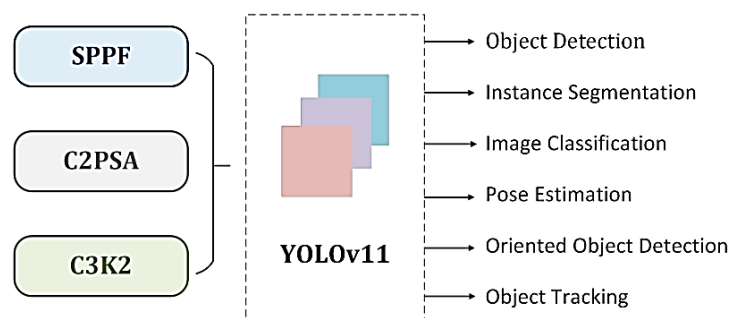


Figura 3.10. Módulos clave en la arquitectura *Yolov11* (Rahima Khanam & Muhammad Hussain, 2024).

C2PSA

C2PSA (Cross Stage Partial with Spatial Attention) permite un enfoque en regiones clave dentro de la imagen, conllevando a una detección más exacta en casos de oclusiones parciales y objetos a escalas pequeñas.

C3k2

C3k2 (Cross Stage Partial with kernel size 2) son utilizados para refinar los mapas de características y hacer más eficiente el proceso debido a su flexibilidad en la configuración de sus parámetros:

- Cuando *c3k = False*, el módulo *C3k2* funciona similarmente al bloque *C2f* ya que, emplea una estructura estándar conocida como *bottleneck*.
- Cuando *c3k = True*, el *bottleneck* se reemplaza por el módulo *C3* que conlleva a un mejoramiento en la extracción de características.

C3k2 disminuye la sobrecarga computacional debido al uso de operaciones convolucionales reducidas para una extracción de características más rápida. Además, es una versión más compacta de *CSP bottleneck*, haciendo la arquitectura más eficiente en términos del número de parámetros.

CBS

El bloque *CBS (Convolution-BatchNorm-Silu)* se centra en la extracción de características relevantes para una detección de objetos exacta, estabiliza y normaliza el flujo de datos e implementa la función de activación *SiLU* que mejora el desempeño del modelo.

Backbone

Esta sección se conforma principalmente por convolucionales y bloques especializados que en conjunto crean mapas de características en diferentes tamaños. Las capas convolucionales son la base del proceso de extracción de características y, se emplean para reducir la resolución de la imagen al mismo tiempo se incrementan el número de canales gradualmente.

La primera innovación en el *backbone* se da al reemplazar los bloques *C2f* originales por los módulos *C3k2* los cuales, son computacionalmente más eficientes ya que emplean convoluciones más pequeñas en comparación con *Yolov8*.

La segunda novedad es la incorporación de una nueva capa llamada *C2PSA (Cross Stage Partial with Spatial Attention)* colocado después de *SPPF (Spatial Pyramid Pooling-Fast)*. La adición de este componente permite que el modelo se enfoque eficientemente en regiones importantes dentro del mapa de características ya que, se mejora la detección de objetos en diferentes escalas y orientaciones.

Neck

Dentro de esta sección, los cambios se dan al reemplazar los bloques *C2f* por *C3k2*, con el fin de hacer más eficiente y veloz el proceso de agregación de características.

Head

Las mejoras elaboradas en esta sección también involucran el empleo de CBS para refinar los mapas de características y C3k2 para hacer más eficiente el proceso de predicción. Estos bloques en conjunto son colocados a diferentes profundidades para contribuir al procesamiento de características a diferentes escalas y así, generar un mapa de características refinado.

c) YOLOv12 (Tian et al., 2025)

A principios del año 2025, se integra una nueva versión a la familia de YOLO. YOLOv12 está basado en su predecesora YOLOv11 y disponible en cinco tamaños: *Nano*, *Small*, *Medium*, *Large* y *XLarge*.

YOLOv12 es la primera versión en proponer e implementar un mecanismo de atención basados en los transformadores de visión denominada como *Area Attention (A2)*. Esta innovación trae como ventaja un amplio campo receptivo con las siguientes particularidades: I) reducción del costo computacional y II) mejor desempeño en la velocidad de inferencia. En la Figura 3.11 se muestra un ejemplo del área de atención. La estrategia del área de atención se caracteriza por dividir de manera simple el mapa de características en l áreas verticales u horizontales.

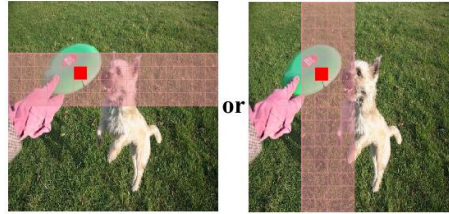


Figura 3.11. Área de atención (Tian et al., 2025).

Es decir, el mapa de características f cuya resolución está dada por (H, W) se divide en l segmentos de tamaño $(\frac{H}{l}, W)$ o $(H, \frac{W}{l})$. Esto elimina las particiones ambiguas y por lo tanto, se logra una velocidad de proceso más rápida. Empíricamente, se asigna $l=4$ lo que reduce el campo receptivo a $\frac{1}{4}$ del tamaño original de f , al mismo tiempo se mantiene un campo receptivo amplio. De modo que, el costo computacional del mecanismo de atención se reduce de $2n^2hd$ a $\frac{1}{2}n^2hd$, donde n es el número de *tokens*, h es el número de cabezales, d es el tamaño del cabezal.

Otra mejora que se menciona ocurre con la creación del módulo residual *R-ELAN*, que se basa en la arquitectura de *Efficient layer aggregation networks (ELAN)*. Este bloque se incorpora solamente en las topologías de gran tamaño de Yolo (*Large and XLarge*) ya que, las pruebas empíricas mostraron problemas de inestabilidad y convergencia en dichas redes aun cuando, se implementan los optimizadores *Adam* o *Adam W*. En la Figura 3.12, se grafica la arquitectura propuesta de *R-ELAN*.

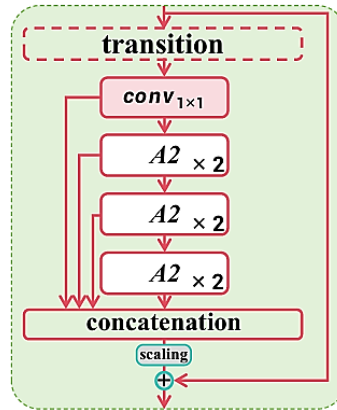


Figura 3.12. Módulo residual *R-ELAN* (Tian et al., 2025).

R-ELAN aplica una capa de transición para ajustar la dimensión de canales y generar un mapa de características único, el cual es procesado por los bloques subsecuentes, seguido por una concatenación y así generar la estructura *bottleneck*. Esta configuración no sólo preserva la capacidad de integración de funciones originales, sino que reduce el costo computacional.

Adicionalmente se realizan las siguientes modificaciones al mecanismo de atención original para tener una mejor compatibilidad con la arquitectura base Yolo: I) ajuste del factor de expansión dimensional del *MLP* de 4.0 a 1.2 (2.0 en los modelos *Nano*, *Small* y *Medium*) para un mejor aprovechamiento de recursos y mejor desempeño, II) incorporación de *nn.Conv2d+BN* para aprovechar al máximo la eficiencia de las operaciones convolucionales y III) implementación de una convolución separable (7x7) llamada *position perceiver*, clave en el área de atención para estimar información posicional.

3.4 Redes especializadas en clasificación

a) ViT (Dosovitskiy et al., 2021)

Un ViT (*Visual Transformer*) o transformador visual es una arquitectura que se utiliza en tareas de reconocimiento de imágenes, a través de una funcionalidad particular llamada “atención”. En la Figura 3.13, se hace una ejemplificación visual del mecanismo de atención en dos imágenes.

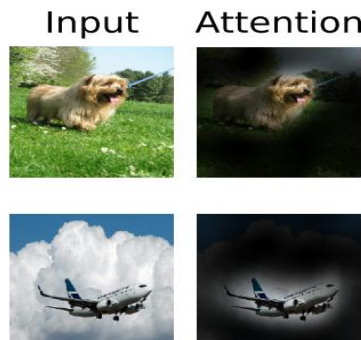


Figura 3.13. Representaciones de atención (Dosovitskiy et al., 2021).

En el caso particular de clasificación de imágenes, el ViT se enfoca en un único objeto de interés en la fotografía, excluyendo la mayor parte que no figura como dato relevante para el problema estudiado.

La manera en que un ViT procesa una imagen para fines de clasificación es el siguiente: la imagen $x \in \mathbb{R}^{H \times W \times C}$ que el ViT recibe como dato de entrada se divide en pequeñas cuadrículas secuenciales sin traslapes llamadas “parches” $x_p \in \mathbb{R}^{N \times (p^2 \times C)}$, donde (H, W) es la resolución de la imagen original, C es el número de canales de la foto, (P, P) es la resolución de cada parche y $N = \frac{HW}{p^2}$ es el número total de parches en la imagen.

Cada parche se transforma en un vector de D dimensiones (Ecuación 3.7), este proceso se conoce como proyección lineal.

$$Z_0 = [X_{class}; X_p^1 E; X_p^2 E; \dots X_p^N E] + E_{pos}, \quad E \in \mathbb{R}^{N \times (p^2 \times C) \times D}, \quad E_{pos} \in \mathbb{R}^{(N+1) \times D} \quad \text{Ecu. 3.7}$$

dónde:

- Z_0 Contiene el *token* de clasificación y los *embeddings* correspondientes a los N parches de la imagen, incluyendo información posicional.
- X_{class} Es un vector entrenable que se incorpora al principio de la secuencia y se utiliza para representar globalmente la imagen.
- X_p^i Parche i -ésimo de la imagen, donde $i=1, \dots, N$. Cada parche contiene los valores de sus píxeles.
- E Matriz de proyección o *embedding*. Convierte cada parche a un vector de dimensión D .
- E_{pos} Añade información sobre la posición espacial que ocupa cada parche dentro de la imagen.

Adicionalmente, se antepone una incrustación destinada como identificador de clase a la proyección lineal de parches ($z_0^0 = X_{class}$), el cual también es transformado en un vector (z_L^0), para representar la imagen Y (Ecuación 3.8).

$$Y = LN(z_L^0) \quad \text{Ecuación 3.8}$$

Enseguida, se establece el número de posición de cada proyección lineal, para ello se utiliza el *standard learnable 1D position embeddings*. Tras el cálculo del posicionamiento este se adjunta a cada proyección, generando así una secuencia de *embedding vectors* que será la entrada al codificador.

Dentro del codificador del transformador, se encuentran alojadas varias capas *Multiheaded Self-Attention (MSA)* y bloques *MLP (Multi Layer Perceptron)* (Ecuaciones 3.9 y 3.10).

$$z'_\ell = MSA(LN(z_{\ell-1})) + z_{\ell-1}, \quad \ell = 1 \dots L \quad \text{Ecuación 3.9}$$

$$z_\ell = MLP(LN(z'_\ell)) + z'_\ell, \quad \ell = 1 \dots L \quad \text{Ecuación 3.10}$$

dónde:

- z'_ℓ Representación intermedia obtenida después de aplicar *MSA* + conexión residual al bloque ℓ .
- z_ℓ Representación de salida del bloque *Transformer* ℓ , después de aplicar *MLP* + conexión residual.
- ℓ Índice del bloque *Transformer* a procesar.
- L Número total de bloques *Transformer Encoder* de la arquitectura.

Los módulos de normalización (*NORM*), se aplican antes de cada una de estas capas y las conexiones residuales a la salida de cada bloque respectivo.

En la salida del codificador del *ViT*, se encuentra un *MLP* empleado como un clasificador. En la Figura 3.14, se muestra de manera genérica la representación de un *ViT*, así como las partes que integran el codificador del transformador.

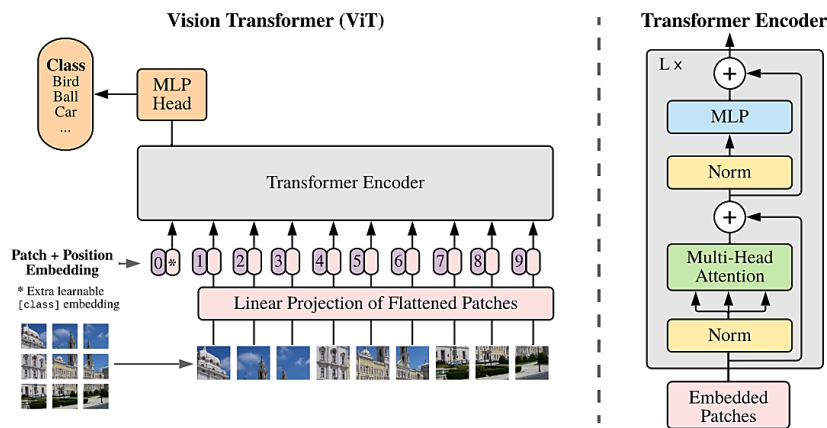


Figura 3.14. Transformador visual (*ViT*) (Dosovitskiy et al., 2021).

El modelo *ViT* interpreta la imagen de entrada como una secuencia de parches (*tokens*), semejante a los transformadores usados en *NLP* (*Natural Language Processing*), donde extensos textos se procesan hasta obtener una serie de *tokens*.

En la literatura se encuentran publicaciones de otras variantes de *ViT*, por ejemplo: I) *MobileSAM* (Chaoning et al., 2023) un modelo enfocado en la segmentación de imágenes y II) *MobileViT* (Sachin & Mohammad, 2022) un modelo híbrido que utiliza operaciones convolucionales basándose en la red *MobileNetv2* (Sandler et al., 2018) en conjunto con módulos de atención basado en *ViT* para tareas de clasificación.

b) MobileViT (Sachin & Mohammad, 2022)

MobileViT es una arquitectura híbrida que implementa operaciones convolucionales basadas en el modelo *MobileNetv2* y, una combinación de módulos de atención basado en *ViT* para efectos de clasificación de imágenes.

Los *ViTs* son modelos orientados para aprender representaciones visuales a través de imágenes de una manera distinta a las redes *CNNs* (Convolutional Neural Networks) ya que, la muestra digital a procesar se divide en una secuencia de parches, los cuales no se encuentran superpuestos unos contra otros. A partir de esta separación los *ViTs* utilizan una serie de múltiples cabezales de atención para generar distintas contextualizaciones de dicha muestra digital.

MobileViT es un bloque que codifica la información local y global, es decir, combina los beneficios tanto de las *CNNs* (sesgos inductivos espaciales y menor sensibilidad al aumento de datos) como de los *ViTs* (ponderación adaptativa de entrada y procesamiento global). En la Figura 3.15 se ilustra: a) un transformador visual estándar; b) la topología del bloque *MobileViT*.

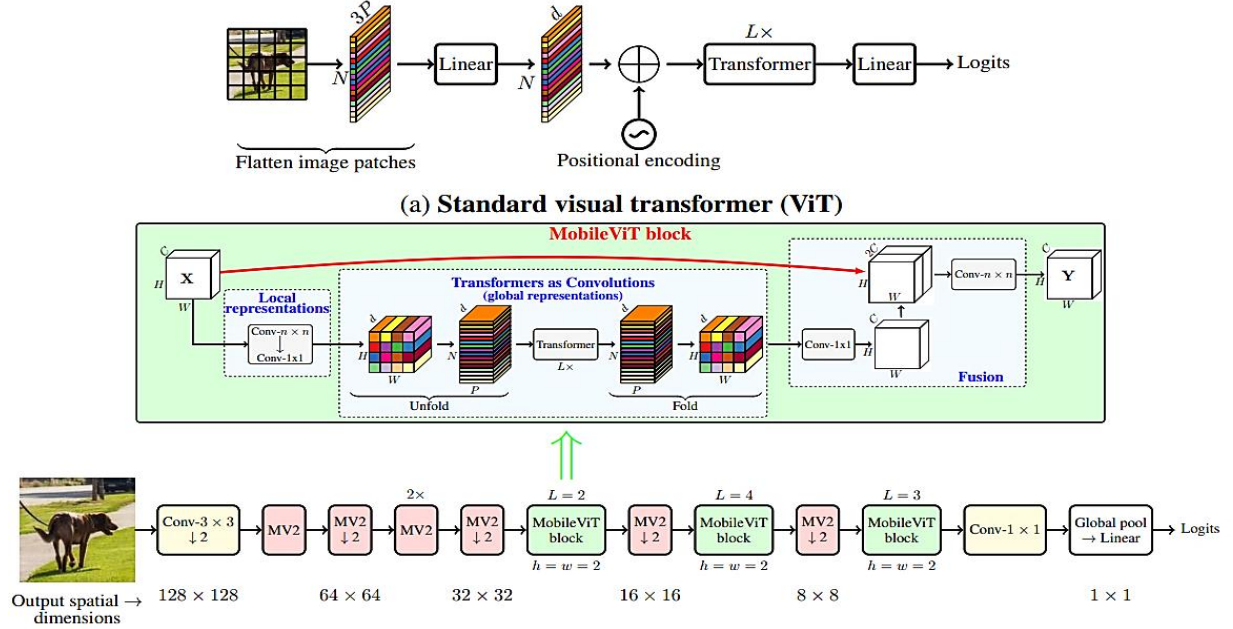


Figura 3.15. Arquitectura *MobileViT* (Sachin & Mohammad, 2022).

La capa inicial de la arquitectura es una convolución estándar de 3×3 ($n = 3$), seguida por cuatro bloques *MV2* (*MobileNetV2*) consecutivos y bloques *MobileViT*, los cuales son tres y están intercalados con los módulos *MV2*. Por un lado, los bloques *MV2* contribuyen a reducir la resolución de los mapas de características. El espacio dimensional de dichos mapas se multiplica por 2; $(h, w) \leq n$. Entonces, $h = w = 2$. La función de activación definida es *Swish*. Por otro lado, los parámetros de los módulos de atención son los siguientes: $L = \{2, 4, 3\}$ y $d = \{96, 120, 144\}$ en sus respectivos niveles espaciales de 32×32 , 16×16 , y 8×8 .

El bloque *MobileViT* recibe un tensor de entrada $X \in \mathbb{R}^{H \times W \times C}$, el cual es procesado por una operación convolucional estándar (*Conv- $n \times n$*) seguido por la capa convolucional *point-wise* (1×1) para producir un nuevo tensor $X_L \in \mathbb{R}^{H \times W \times d}$. La primera capa se enfoca en codificar la información espacial local mientras que, la capa *point-wise* se encarga de generar las altas proyecciones dimensionales espaciales (donde $d > C$) a partir de la combinación de los canales de entrada-ya que, la idea es modelar dependencias no locales de largo alcance.

Para que *MobileViT* aprenda las representaciones globales con un sesgo espacial inductivo, el tensor X_L se procesa para obtener N non-overlapping flattened patches $X_U \in \mathbb{R}^{P \times N \times d}$, donde $P = wh$, $N = \frac{HW}{P}$ determina el número de parches; $h \leq n$ y $w \leq n$ corresponden al

alto y ancho de cada parche respectivo. Las relaciones de cada parche $p \in \{1, \dots, P\}$, son codificados al emplear transformadores para obtener $X_G \in \mathbb{R}^{P \times N \times d}$:

$$X_G(p) = \text{Transformer}(X_U(p)), 1 \leq p \leq P$$

Para que *MobileViT* no pierda el orden de los parches ni el orden espacial de píxeles dentro de cada parche, se procesa $X_G \in \mathbb{R}^{P \times N \times d}$ hasta obtener $X_F \in \mathbb{R}^{H \times W \times d}$. X_F se proyecta al espacio de baja dimensión C al utilizar la capa convolucional *point-wise* combinado con X mediante una operación de concatenación. Otra capa convolucional $n \times n$ se emplea para fusionar las características de X_U y X_G puesto que, $X_U(p)$ codifica información local de región $n \times n$ usando convoluciones y $X_G(p)$ codifica información global tras la secuencia de cada posición de los parches P , donde cada píxel en X_G codifica información de todos los píxeles en X , tal y como se ilustra en la Figura 3.16.

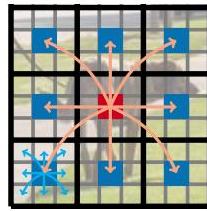


Figura 3.16. Multiconexión entre píxeles en el bloque *MobileViT* (Sachin & Mohammad, 2022).

En este ejemplo, las celdas en color negro representan los parches y las cuadrículas en color gris son los píxeles. Los cuadrados azules alojan información de los píxeles vecinos usando operaciones convolucionales mientras que, los transformadores de visión al enfocarse en un píxel (zona roja) sin importar su ubicación se obtiene información global de todos los píxeles que conforma la imagen.

3.5 Redes enfocadas en segmentación

a) *MobileSAM* (Chaoning et al., 2023)

Es un modelo de segmentación de imágenes creado principalmente para ser implementado en dispositivos que cuenten con pocos recursos de hardware. El desarrollo de esta arquitectura se basa en el modelo original *SAM* (*Segment Anything Model*) (Kirillov et al., 2023). En la Figura 3.17, se ilustra de manera general la arquitectura del modelo *SAM*.

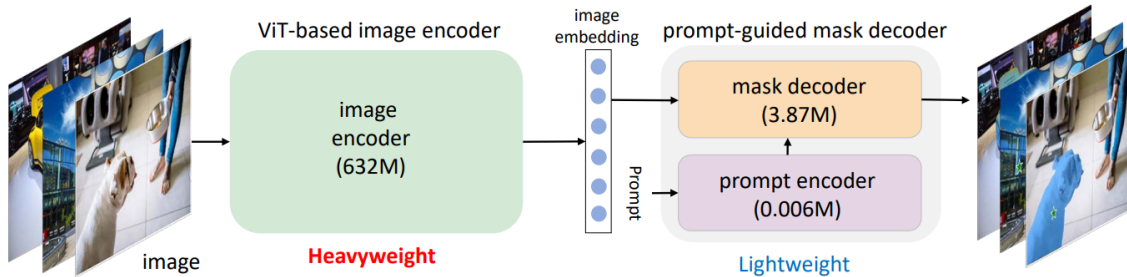


Figura 3.17. Esquema del modelo *SAM* (Chaoning et al., 2023).

El codificador de imagen (*image encoder*) está basado en un *ViT-H (large version)* con 632 millones de parámetros, volviéndolo totalmente incompatible para ser implementada en dispositivos móviles.

Para ello, se busca remplazar el extenso codificador de imágenes original (*ViT-based image encoder*) por una versión más reducida por medio del proceso llamado “destilación del conocimiento acoplado”. En la Figura 3.18, se ilustran dos enfoques de dicho proceso.

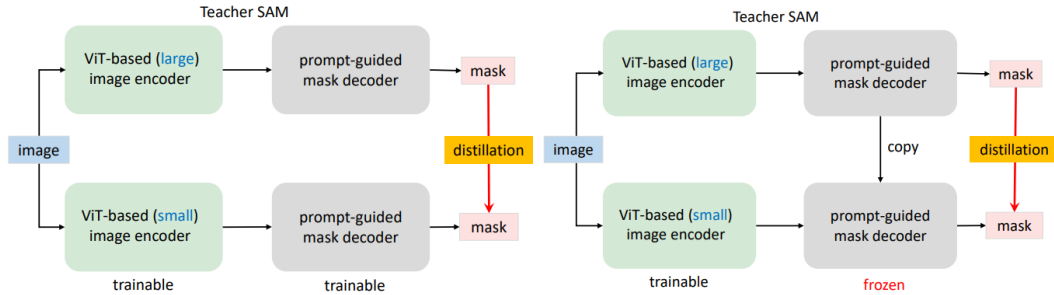


Figura 3.18. Representación de dos tipos de destilación del conocimiento acoplado del modelo SAM (Chaoning *et al.*, 2023).

A la izquierda, se ilustra el esquema de destilación totalmente acoplado mientras que, a la derecha se encuentra la destilación semi acoplada. En ambos casos, el *ViT-based (large)* tiene el rol de “maestro” y *ViT-based (small)* como “estudiante”. Además, se emplean las máscaras decodificadoras como un *prompt guiado*. A Este proceso de reentrenamiento para transferir el conocimiento de maestro a estudiante se le conoce como destilación del conocimiento (*knowledge distillation KD*).

La metodología establecida busca mejorar la transferencia de los pesos del codificador *ViT* a un nuevo codificador más pequeño. En la Figura 3.19, se muestra el esquema propuesto llamado “destilación del conocimiento desacoplado”

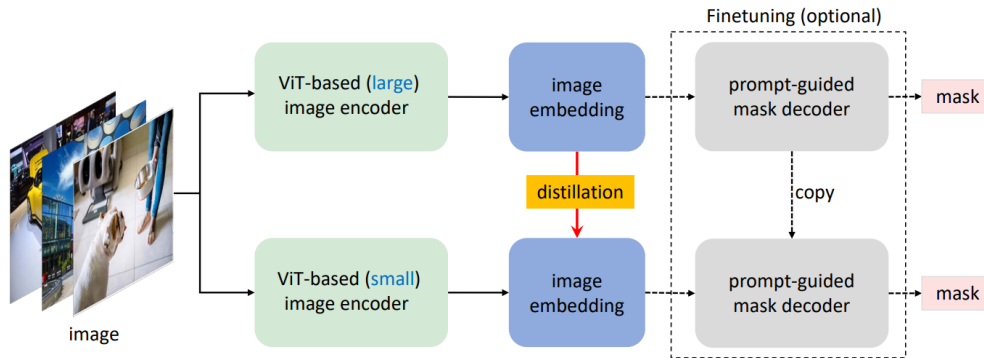


Figura 3.19. Destilación desacoplada (Chaoning *et al.*, 2023).

El término desacoplado hace referencia a que no se contempla el uso de la máscara decodificadora ya que, la codificación de la imagen generada por el codificador de imágenes del estudiante es lo suficientemente cercana al del codificador maestro original, lo que hace que el ajuste fino del decodificador combinado en la segunda etapa sea opcional. Se utiliza

sólo el 1% del *dataset SA-1B* (Kirillov *et al.*, 2023) para el proceso de entrenamiento y así generar el modelo de segmentación propuesto.

En la Tabla 3.2, se hace la comparación del número de parámetros y tiempo de inferencia de los modelos *SAM* y *MobileSAM*. La velocidad de inferencia se calcula solamente con una *GPU*.

Tabla 3.2. Comparación de parámetros y rapidez de inferencia (Chaoning *et al.*, 2023).

	Original SAM	MobileSAM
Parameters	632M	5.78M
Speed	452ms	8ms

En la Figura 3.20, se ilustra una equiparación de los resultados obtenidos con el modelo *MobileSAM* y *Original SAM*, empleando como *prompt* de entrada un *BB*.



Figura 3.20. Predicción de una máscara de píxeles, delimitando al objeto con un *BB* (Chaoning *et al.*, 2023).

b) SAM2 (Ravi *et al.*, 2024)

Es un modelo enfocado en resolver problemas de segmentación tanto en imágenes como en videos (nueva función), empleando los cuadros delimitadores (*BB*), máscaras de píxeles y puntos como *prompt* de entrada para refinar la extensión espacial (área) del objeto a segmentar. Dentro de las mejoras, se encuentra la incorporación de módulos de memoria que almacena información del objeto y las iteraciones anteriores, lo que le permite generar predicciones de máscara a lo largo del video y también corregirlas de manera efectiva según el contexto de memoria almacenado del objeto. La Figura 3.21, muestra la arquitectura de la red *SAM2*.

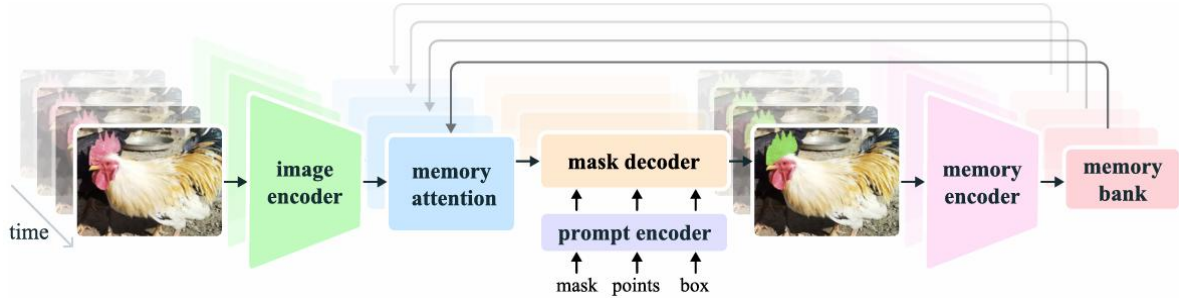


Figura 3.21. Topología de SAM2 (Ravi et al., 2024).

El rol de la memoria de atención (*memory attention*) es condicionar las características del *frame* actual sobre las características y predicciones de los *frames* pasados, así como cualquier nuevo *prompt*.

La memoria codificadora (*Memory encoder*) define un espacio de memoria al submuestrear la máscara de salida mediante un módulo convolucional y sumando elemento por elemento con la incrustación de fotogramas no condicionados del codificador de imágenes, seguido de capas convolucionales ligeras para fusionar la información.

El banco de memoria (*memory bank*) retiene información de predicciones pasadas del objetivo en el video, gestionada en una cola *FIFO* de N frames e información de los *prompts* en una cola *FIFO* de M prompted frames.

3.6 Configuraciones específicas

a) *Depth Wise Separable Convolution (DWSC)*

Es una agrupación de operaciones convolucionales conformada por dos procesos. En la primera etapa (*Depth-wise convolution*), por lo general se emplea una convolución usando un filtro de $c \times c$; $c \geq 3$, por cada uno de los canales M que integra un mapa de características F . Posteriormente, en la segunda etapa (*Point-wise convolution*), se efectúa una segunda convolución utilizando un filtro de $c \times c$; $c = 1$, partiendo del conjunto de filtros creados previamente. En la Figura 3.22 se ejemplifican: a) *Depth-wise Separable Convolution*; y b) convolución estándar.

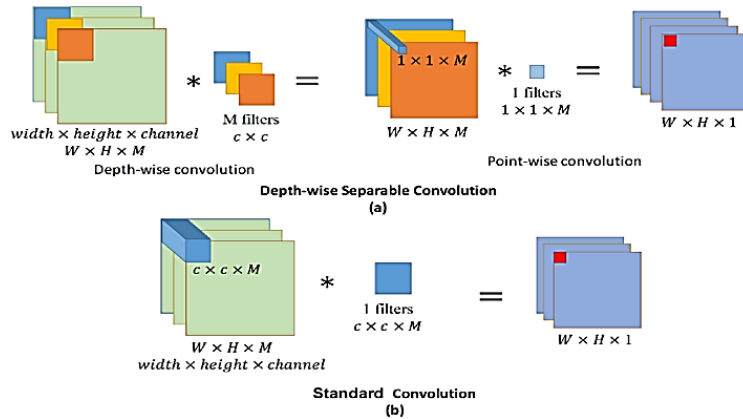


Figura 3.22. Comparativa de operaciones convolucionales (Hüseyin et al., 2022).

La mayoría de modelos ligeros emplean dicha configuración para disminuir el número de operaciones de punto flotante (*FLOPs*) que se miden en segundos, preservando la información de las características sin cambios (Wang et al, 2023).

La red *MobileViT* implementa *DWSC* basado en *MobileNetV2*, para reducir las dimensiones de los mapas de características, minimizando el costo computacional (Sachin & Mohammad, 2022).

Por ejemplo, los *FLOPs* requeridos para ejecutar una convolución regular con un tamaño del filtro $c \times c$; $c \geq 3$. Cada filtro de dimensiones espaciales $c \times c$ se extiende a través de los M canales de entrada. Al emplearse M filtros para generar M mapas de características de salida. La Ecuación 3.11 expresa el costo computacional.

$$FLOPs_{Conv} = H \times W \times c^2 \times M^2 \quad \text{Ecuación 3.11}$$

En cuanto a los *FLOPs* necesarios para cuantificar una *Deep Convolution*, se determina mediante la Ecuación 3.12. En este caso la profundidad del filtro no crece proporcionalmente al número de canales a calcular ya que, el filtro procesa cada canal por separado.

$$FLOPs_{DWConv} = H \times W \times c^2 \times M \quad \text{Ecuación 3.12}$$

b) *CBAM* (Woo et al., 2018)

CBAM (*Convolution Block Attention Module*), es un método propuesto para mejorar el desempeño de una red neuronal convolucional orientada para tareas de clasificación y detección a partir de operaciones convolucionales. *CBAM* enfatiza las características relevantes partiendo de dos módulos, los cuales se encargan de inferir atención con enfoques diferentes. En la Figura 3.23 se muestra la topología de *CBAM*.

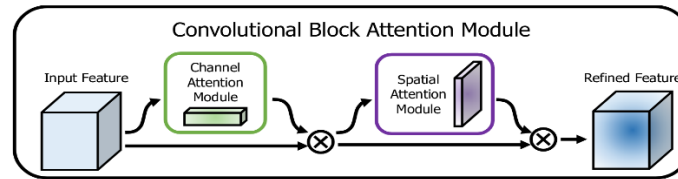


Figura 3.23. Esquema general de *CBAM* (Woo et al., 2018).

Sea $F \in \mathbb{R}^{C \times H \times W}$ el mapa de características intermedio empleado como entrada al *CBAM*, secuencialmente se genera un mapa de canal de atención 1D $M_c \in \mathbb{R}^{C \times 1 \times 1}$ y un mapa de atención espacial 2D $M_s \in \mathbb{R}^{1 \times H \times W}$. Las Ecuaciones 3.13 y 3.14 expresan el proceso de *CBAM*:

$$F' = M_c(F) \otimes F, \quad \text{Ecuación 3.13}$$

$$F'' = M_s(F') \otimes F', \quad \text{Ecuación 3.14}$$

donde $\otimes$, denota la multiplicación *element-wise*. Los valores del canal de atención son propagados a lo largo de la dimensión espacial y viceversa. F'' es la salida (mapa de

características refinado). La Figura 3.24 ilustra el proceso del cálculo de cada mapa de atención.

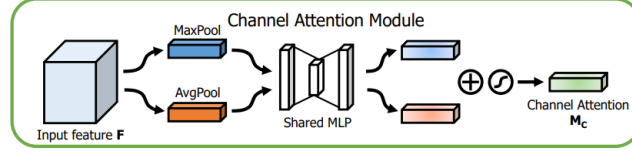


Figura 3.24. Diagrama del módulo de atención de canal (Woo et al., 2018).

Dentro del módulo de atención de canales, se agrega información espacial a F al emplear las operaciones *average-pooling* y *max-pooling*, dando origen a dos descriptores espaciales contextuales F_{avg} y F_{max} , (características de *average-pooling* y *max-pooled* respectivamente). Enseguida, ambos descriptores son procesados por el perceptrón multicapa (*shared MLP*) compuesto por una capa oculta para definir el mapa de atención de canales $M_c \in \mathbb{R}^{C \times 1 \times 1}$. Para reducir el número de parámetros, el tamaño de la capa de activación se define como $\mathbb{R}^{C \frac{1}{r \times H \times W}}$, donde r es la tasa de reducción. A la salida del perceptrón multicapa, se unen los vectores descriptivos utilizando *element-wise summation*. Formalmente (Ecuación 3.15), el módulo de atención de canales se define:

$$M_c(F) = \sigma \left(MLP(AvgPool(F)) + MLP(MaxPool(F)) \right) \quad \text{Ecuación 3.15}$$

$$M_c(F) = \sigma \left(W_1 \left(W_0(F_{avg}^c) \right) + W_1 \left(W_0(F_{max}^c) \right) \right)$$

Donde σ denota la función *sigmoid*; W_0 , y W_1 son pesos del *MLP*.

En el módulo de atención espacial, al mapa F' se aplican las operaciones de *average-pooling* y *max-pooling* para generar un descriptor de características. Posteriormente, ambas capas son concatenadas y procesadas por una convolución estándar, definiendo así un mapa de atención espacial $2D$, el cual codifica dónde enfatizar o suprimir la información. La Figura 3.25, esquematiza el bloque de atención espacial.

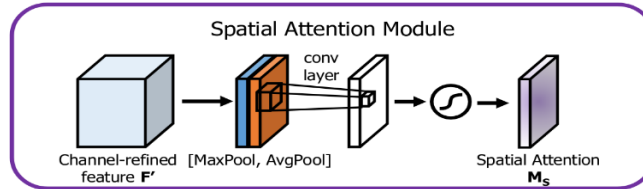


Figura 3.25. Diagrama del módulo de atención espacial (Woo et al., 2018).

La Ecuación 3.16, define el módulo de atención espacial:

$$M_s(F) = \sigma \left(f^{7 \times 7}([AvgPool(F); MaxPool(F)]) \right) \quad \text{Ecuación 3.16}$$

$$M_s(F) = \sigma \left([F_{avg}^s; F_{max}^s] \right)$$

Donde σ denota la función *sigmoid* y $f^{7 \times 7}$ representa la operación convolucional usando un filtro de tamaño 7×7 .

El módulo de atención de canal se centra en calcular qué características son relevantes en cada canal, sin contemplar la información espacial, mientras que el bloque de atención espacial se enfoca en calcular dónde se encuentran las propiedades más importantes, lo cual complementa al módulo de atención de canal.

Puesto que la configuración del *CBAM* no tiene un impacto significativo en términos de costo computacional y es de propósito general, la integración con otras arquitecturas basadas en operaciones convolucionales es fácil de implementar, contribuyendo al aprendizaje de la red para discriminar la información relevante y no importante.

3.7 Funciones de regresión para el cálculo de instancias delimitadoras

A continuación, se presentan tres funciones de regresión orientadas para tareas de detección. Cada método propuesto por (Wang *et al.*, 2023), se enfoca en estimar las coordenadas de los cuadros delimitadores que encasillan al objeto de interés en la imagen.

a) Función de regresión *WIoUv1*

La función *WIoUv1* (*Wise Intersection over Union*) implementa un enfoque donde, la distancia de centroides actúa como mecanismo de atención de modo que, cuando el *Bounding Box gt* y *Bounding Box predicted* se traslapan dentro de un determinado rango, la penalización estimada hace que el modelo obtenga una mejor inferencia. Las ecuaciones 3.17-3.19 definen *WIoUv1*.

$$L_{WIoUv1} = R_{WIoU} \times L_{IoU} \quad \text{Ecuación 3.17}$$

$$R_{WIoU} = \exp\left(\frac{(b_{cx}^{gt} - b_{cx})^2 + (b_{cy}^{gt} - b_{cy})^2}{((c_w)^2 + (c_h)^2)}\right) \quad \text{Ecuación 3.18}$$

$$L_{IoU} = 1 - IoU \quad \text{Ecuación 3.19}$$

b) Función de regresión *WIoUv2*

Es una extensión de *WIoUv1*, que implementa un coeficiente focal monótono L^*_{IoU} que reduce eficazmente el peso de los ejemplares simples de modo que, el modelo tiende a enfatizar las instancias difíciles lo que contribuye a una mejora en la clasificación (Tong *et al.*, 2023). En la ecuación 3.20 se define la función *WIoUv2*.

$$L_{WIoUv2} = \left(\frac{L^*_{IoU}}{L_{IoU}}\right)^\gamma \times L_{WIoUv1}; \gamma > 0 \quad \text{Ecuación 3.20}$$

Durante el entrenamiento iterativo L^*_{IoU} decrece al igual que L_{IoU} , dando como resultado una convergencia más lenta. Por lo tanto, se emplea el promedio de L_{IoU} para normalizar L^*_{IoU} (Wang G. *et al.*, 2023).

c) Función de regresión *WIoUv3*

Es una versión mejorada que involucra a *WIoUv1* que incorpora un coeficiente focal no monótono β . *WIoUv3* se centra en cuantificar la calidad de los *Anchor Boxes (AB)*, contribuyendo a que el modelo se enfoque más en los *AB* de calidad ordinaria y, por ende, un mejor desempeño para localizar los objetos. En las ecuaciones 3.21 – 3.23 se define la función *WIoUv3*.

$$L_{WIoUv3} = r \times L_{WIoUv1} \quad \text{Ecuación 3.21}$$

$$r = \frac{\beta}{\delta \alpha^{\beta-\delta}} \quad \text{Ecuación 3.22}$$

$$\beta = \frac{L^*_{IoU}}{L_{IoU}} \in [0, +\infty) \quad \text{Ecuación 3.23}$$

Un valor pequeño de β indica una alta calidad del AB y, se le asigna un r (ganancia del gradiente) más pequeño, lo que reduce la ponderación de los AB de calidad alta en el cálculo de la función de pérdida. Un valor alto de β señala una calidad baja del AB entonces, se le asigna un valor pequeño de r , lo cual reduce alteraciones negativas a la gradiente causados por los AB de baja calidad (Wang G. *et al.*, 2023).

En la Figura 3.26, se ilustra el mapeo de valores atípicos β , la ganancia del gradiente r , que son controlados por los hiperparámetros α y δ , donde δ hace que $r = 1$ cuando $\beta = \delta$. El AB alcanza la mayor ganancia del gradiente cuando su grado atípico satisface $\beta = C$ (C , es una constante) (Tong *et al.*, 2023).

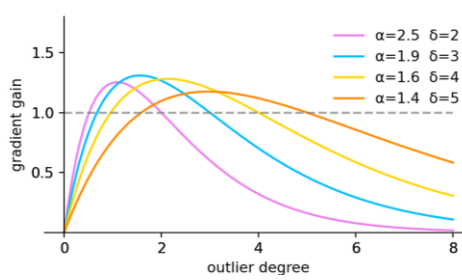


Figura 3.26. The mapping of outlier degree (Tong *et al.*, 2023).

WIoUv3 implementa una estrategia de ganancia del gradiente que optimiza dinámicamente la alta y baja calidad de los AB, haciendo que el modelo se enfoque en la calidad promedio de los AB y por consiguiente, se mejore el desempeño general del modelo.

La justificación del remplazo de la función original *CIoU* (*Complete Intersection over Union*) del modelo *Yolov11n* por la versión *WIoUv3* se aborda en el capítulo cinco (Método de solución).

3.8 Métricas de evaluación

Durante la experimentación se recurre al uso de indicadores estándares y/o herramientas que son empleados para analizar y medir el desempeño del modelo, esto con el fin de comparar los resultados obtenidos contra otros puntajes reportados en el estado del arte y, realizar un análisis detallado para llegar a una conclusión que permita tomar las decisiones pertinentes del producto en desarrollo.

a) Inspección cualitativa (Grandi et al., 2020)

La matriz de confusión (Tabla 3.3) es una herramienta que permite hacer una inspección de manera visual para conocer el estatus del modelo. En esta tabla se comparan los datos predichos por el modelo con respecto a la información definida por el científico de datos. Los valores que se encuentran alojados en las celdas son: *TP* (*True Positive*), *TN* (*True Negative*), *FP* (*False Positive*) y *FN* (*False Negative*) los cuales, son datos numéricos y se emplean en las métricas de evaluación para cuantificar un valor final.

Tabla 3.3. Matriz de confusión.

Modelo		Científico	
		True	False
	True	TP	FP
	False	FN	TN

Los registros en color verde indican que las predicciones que realiza el modelo coinciden con los datos del científico. Por el contrario, la información en color rojo señala que existen inconsistencias por parte del modelo.

Esta revisión también aplica cuando se hace la comparación entre dos máscaras de píxeles o dos cuadros delimitadores que figuran en una imagen.

En la Figura 3.27, se aprecian dos máscaras (clase hexágono) y un fondo blanco (clase fondo), la máscara de contorno color rojo es inferida por el modelo mientras que, la máscara de contorno color negro es establecida por el científico de datos.

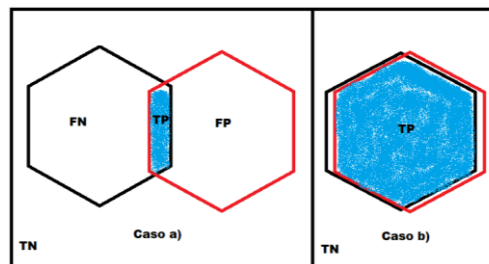


Figura 3.27. Comparación de máscaras de píxeles.

En el caso a) se observa que la máscara roja escasamente logra cubrir uno de los seis lados de la figura negra lo cual significa que, el modelo no es capaz de clasificar más píxeles que pertenecen a la clase hexágono (color azul), confunde el resto del hexágono como si fuera el fondo dicho en otras palabras la predicción es inadecuada. En el caso b) se aprecia todo lo contrario, ahora la máscara roja cubre casi en su totalidad a la máscara hecha por el científico, por lo tanto, el modelo genera máscaras de píxeles asemejándose a la forma de un hexágono.

Semejante al caso anterior en la Figura 3.28, se ilustran dos cuadros delimitadores el tono rojo calculado por el modelo y el tono oscuro definido por el científico de datos el cual, encasilla una estrella.

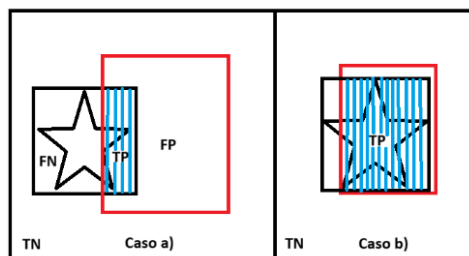


Figura 3.28. Comparación de cajas delimitadoras.

En el caso a), se observa que el modelo no realiza una correcta estimación de las coordenadas del rectángulo, lo que ocasiona una imprecisión importante para ubicar el espacio donde yace el objeto. Sin embargo, en el caso b) la predicción del modelo es más precisa debido a que, las proporciones del *BB* visualmente se asemejan a las dimensiones definidas manualmente.

b) Cálculo cuantitativo

A partir de los valores numéricos de los casos expresados en la matriz de confusión (Tabla 3.3), se definen siete métricas de evaluación que cuantifican un valor porcentual definido en $[0,1]$ donde, un dígito cercano a 1 señala un óptimo desempeño, una cantidad aproximadamente a 0.5 y por debajo de este indica un funcionamiento pobre del modelo bajo estudio.

En la Tabla 3.4 (pp. 43 - 44) se explican cada uno de los indicadores categorizados para tareas de detección y/o segmentación, donde: TP, TN, FP y $FN \in \mathbb{N}$ corresponden al conteo de objetos presentes en la imagen. Mientras que, $\sum TP, \sum TN, \sum FP$ y $\sum FN \in \mathbb{N}$ hace referencia al conteo de píxel por píxel que conforman diferentes regiones de interés en la imagen.

Tabla 3.4. Métricas de evaluación.

Métrica	Dominio	Descripción	Ecuación
<i>Accuracy</i>	Detección, Segmentación	Es el porcentaje global de objetos o píxeles clasificados correctamente en la imagen.	$Acc = \left(\frac{\sum TP + \sum TN}{\sum TP + \sum TN + \sum FP + \sum FN} \right)$ $Acc = \left(\frac{TP + TN}{TP + TN + FP + FN} \right)$ $Acc \in \mathbb{R}; 0 \leq Acc \leq 1$
<i>Precision</i>	Detección, Segmentación	Es el porcentaje de aciertos de los píxeles u objetos que son correctamente inferidos con respecto a la máscara de píxeles o coordenadas definidos en el <i>GT</i> .	$Pr = \left(\frac{\sum TP}{\sum TP + \sum FP} \right)$ $Pr = \left(\frac{TP}{TP + FP} \right)$ $Pr \in \mathbb{R}; 0 \leq Pr \leq 1$
<i>Recall</i>	Detección, Segmentación	Es el porcentaje de los objetos o píxeles que se encuentran dentro	$Re = \left(\frac{\sum TP}{\sum TP + \sum FN} \right)$

Tabla 3.4. Métricas de evaluación.

Métrica	Dominio	Descripción	Ecuación
		del perímetro del cuadro delimitador o máscara predicha con respecto al <i>GT</i> . Un valor alto de <i>Recall</i> señala un menor escape de objetos o píxeles (Abulizi A et al, 2025)	$Re = \left(\frac{TP}{TP + FN} \right)$ $Re \in \mathbb{R}; 0 \leq Re \leq 1$
<i>F1</i>	Detección, Segmentación	<i>F1</i> mide el balance entre la <i>precision</i> y <i>recall</i> . Si el valor de <i>F1</i> es alto significa que hay un equilibrio favorable. Un valor bajo de <i>F1</i> indica que, existe un gran desbalance entre dichos indicadores (Gad, 2020).	$F1 = \frac{(2)(Pr)(Re)}{(Pr) + (Re)}$ $F1 \in \mathbb{R}; 0 \leq F1 \leq 1$
<i>Accuracy Precision (AP)</i>	Detección	Es el valor del área bajo la curva (AUC) de <i>precision-recall</i> que representa el promedio de todas las precisiones (<i>Pr</i>) (Gad, 2020).	$AP = \sum_{k=0}^{k=n-1} [Re_k - Re_{k+1}] (Pr_k)$ $Ap \in \mathbb{R}; 0 \leq Ap \leq 1$
<i>Mean Average Precision (mAP)</i>	Detección	Es la media ponderada de <i>precision</i> en cada umbral y se define al sumar el <i>AP</i> de cada clase sobre el número total de clases (Gad, 2020).	$mAP = \frac{1}{N} \sum_{i=1}^N AP_i$ $mAP \in \mathbb{R}; 0 \leq mAP \leq 1$
<i>IoU (Intersection over Union); Jaccard Index</i>	Segmentación	Es el porcentaje del área de intersección entre la máscara predicha con respecto a la máscara establecida en el <i>GT</i> .	$IoU = \left(\frac{\sum TP}{\sum TP + \sum FN + \sum FP} \right)$ $IoU \in \mathbb{R}; 0 \leq IoU \leq 1$

CAPÍTULO 4

BANCO DE IMÁGENES

El capítulo detalla la información de diez conjuntos de imágenes provenientes de la plataforma (*RoboFlow*, s.f), también se especifican las características de las imágenes recolectadas a través de diversos sitios webs.

4.1 Bancos de imágenes provenientes de *RoboFlow*

RoboFlow es un repositorio gratuito que gestiona una amplia variedad de bancos de imágenes para visión artificial. Con el fin de garantizar la representatividad del fenómeno bajo estudio, se establecieron criterios estrictos de selección y filtrado del material visual. De este modo, las muestras seleccionadas ilustran escenarios complejos en entornos agrícolas no controlados (campo abierto).

a) *Dataset A4*

El conjunto de datos comprende un total de 906 muestras que se distribuyen en tres particiones: 634 imágenes para el entrenamiento, 181 para validación y 91 de prueba. Cada imagen a color posee una resolución fija de 416×416 píxeles.

Algunas de las características visuales de las fotos son: I) tomas efectuadas en ambientes no controlados, II) presencia de iluminación variable, III) las hojas del jitomate que figuran en la imagen son del tipo papa y sierra, IV) alteraciones en el color, textura y morfología de las láminas foliares. La Figura 4.1 ilustra dichas descripciones.



Figura 4.1. Capturas digitales del banco de imágenes A4 (*Roboflow*, s.f).

La ventaja de este conjunto radica en la definición y nitidez óptica de sus imágenes. No obstante, es necesario señalar dos limitaciones estructurales en su diseño: I) la inclusión de la mano humana en el fondo de la escena introduce un artefacto en el fondo y II) la presencia de una única instancia por imagen restringe el etiquetado a una sola caja delimitadora (*Bounding Box*). Esto último simplifica la tarea de localización y reduce la complejidad frente a escenarios donde existen múltiples objetivos.

b) *Test 7 (T7)*

El conjunto de datos comprende un total de 542 imágenes distribuidas en dos particiones: entrenamiento con 436 muestras y validación con 106 instancias. Los ejemplares capturan grandes conglomerados de hojas del tipo sierra y papa de la planta del jitomate. Las láminas foliares muestran variabilidad fenotípica caracterizada por alteraciones en la geometría de sus contornos y coloración. Adicionalmente, el *dataset* incluye variaciones de escala y en la perspectiva de las tomas, en escenarios agrícolas complejos. La Figura 4.2 destaca algunos ejemplares del contenido.



Figura 4.2. Capturas digitales del banco de imágenes T7 (Roboflow, s.f.).

c) *Plant Computer Vision Project (PCVP)*

El conjunto alberga un total de 380 imágenes las cuales se reparten en tres subconjuntos: entrenamiento (254), validación (89) y prueba (37). Las capturas digitales se centran en agrupaciones foliares densas de jitomate de la variedad tipo sierra. Entre las características visuales más destacadas se encuentran: I) diversidad en los ángulos de captura, II) variabilidad en la escala de las hojas, III) coexistencia de tejido foliar sano y patológico (con alteraciones tanto en color como en forma), IV) fluctuaciones en las condiciones de iluminación y V) presencia de oclusiones. La Figura 4.3 expone ejemplos de estas particularidades.



Figura 4.3. Capturas digitales del banco de imágenes PCVP (Roboflow, s.f.).

Las ventajas de este banco radican en la calidad de la imagen y en la inclusión de múltiples instancias por imagen. Esta densidad y variabilidad permite evaluar la capacidad del modelo en escenarios donde yacen múltiples objetivos.

d) *Tomato Test (TT)*

Contiene un total de 94 muestras y 982 instancias delimitadoras. Este *dataset* no contiene divisiones para las etapas de entrenamiento, validación y prueba; es decir, todas las imágenes se alojan en una carpeta. La resolución que tiene cada ejemplar a color es de 416×416 píxeles.

Las características que tienen las imágenes son: I) tomas realizadas en entornos tanto de exteriores como de interiores, II) variabilidad en los puntos de vista y ángulos de perspectiva, III) las hojas de jitomate corresponden al tipo sierra y IV) las láminas foliares presentan variaciones en color, forma y textura. En la Figura 4.4 se muestran ejemplos de algunos ejemplares.



Figura 4.4. Muestras digitales del banco de imágenes *Tomato Test* (Roboflow, s.f).

Las ventajas de este conjunto son su nitidez óptica y la inclusión de múltiples instancias anotadas por imagen, lo que incrementa el número de etiquetas para el entrenamiento del modelo.

e) Answer Computer Vision Project (ACVP)

El conjunto está constituido por 27 imágenes que contiene un total de 356 instancias anotadas mediante cajas delimitadoras. Las características de cada imagen son: I) capturas realizadas en entornos agrícolas complejos, II) la perspectiva focal está orientada hacia el plano foliar, III) abundantes láminas de la variedad tipo regular y, IV) predominan las hojas de aspecto sano (sin anomalías). Las dimensiones espaciales más frecuentes en las imágenes son de 3,024×4,032 y 4,032×3,024 píxeles. En la Figura 4.5 se ilustran algunas de estas muestras.



Figura 4.5. Tomas de la agrupación de imágenes *Answer Computer Vision Project* (Roboflow, s.f).

Las ventajas de este conjunto son su nitidez óptica y la inclusión de múltiples instancias por imagen.

f) Detecting Calciumn (DC)

La agrupación contiene 133 capturas de hojas de la planta del jitomate del tipo sierra y papa, que presentan síntomas asociados a la deficiencia de calcio. Con base en el nivel de afectación, las muestras se categorizan en cuatro clases que representan distintos estadios de severidad. Visualmente, las hojas presentan distintas variaciones cromáticas que abarcan tonalidades amarillas, crema y algunos degradados. La dimensión espacial promedio es de 3,456×2,304 píxeles con fluctuaciones menores en el tamaño debido al

proceso de segmentación o captura original. En la Figura 4.6, se ejemplifican algunas de las muestras.



Figura 4.6. Fotos de la agrupación de imágenes *Detecting Calciumn* (Roboflow, s.f).

Las ventajas del conjunto son: I) alta definición y nitidez de las imágenes, lo cual permite la correcta preservación y visualización de estructuras como las nervaduras y los lóbulos sobre la lámina foliar, y II) amplia variabilidad reflejada en las dimensiones, la morfología y los atributos cromáticos de las hojas.

g) NAIRS Computer Vision Project (NCVP)

La colección comprende 318 imágenes etiquetadas en dos clases: infectado y saludable. Contiene tomas de hojas de jitomate del tipo sierra cuya resolución es de 640×640 píxeles a color. En la mayoría de los ejemplares se observan láminas foliares de color verde uniforme. No obstante, las regiones de interés focalizan estructuras foliares con anomalías fenotípicas complejas, incluyendo variaciones de coloración, distorsiones morfológicas en el contorno y discontinuidades en la superficie debido a rupturas parciales y perforaciones, etc. También, se aprecian cambios de iluminación y presencia de sombras sobre la superficie principal del objeto de interés.

La distribución del conjunto se divide en tres particiones independientes: entrenamiento (294), validación (12) y pruebas (12). Con el objetivo de expandir la variabilidad de la agrupación de ejemplares y prevenir el sobreajuste del modelo, se aplicaron técnicas de aumentación de datos exclusivamente sobre el subconjunto de entrenamiento, garantizando así la integridad de las fases de validación y prueba. Las transformaciones geométricas y fotométricas implementadas incluyeron: efecto espejo en horizontal y vertical, rotaciones de 90° en el sentido horario y anti horario, cambios en la perspectiva, ajustes en la iluminación (-15% y +15%), saturación (-25% y +25%), exposición (-10% y +10%), desenfoque (2.5px) y tonalidad (-15° y +15°). En la Figura 4.7, se ilustran algunas muestras.



Figura 4.7. Tomas del conjunto *NAIRS Computer Vision Project* (Roboflow, s.f).

El conjunto posee imágenes de buena calidad y asegura la correcta preservación de los descriptores visuales.

h) Tomato Computer Vision Project (TCVP)

El conjunto aloja un total de 1,482 capturas clasificadas en cinco categorías: hoja enferma, flor, hoja, jitomate verde y jitomate rojo. La resolución espacial de cada ejemplar es de 640×640 píxeles.

Las características principales del conjunto son las siguientes: I) hojas de morfología del tipo sierra y papa, II) variedad en escalas y perspectivas de captura, III) diferentes grados de oclusión causados por hojas, frutos, ramas y flores, IV) alteraciones en la iluminación y presencia de diferentes anomalías en la lámina foliar modificando la coloración y forma. La agrupación de las imágenes se divide en tres particiones independientes: entrenamiento (1,022), validación (294) y pruebas (166). En la Figura 4.8 se muestran ejemplos representativos de las imágenes.



Figura 4.8. Imágenes del conjunto *Tomato Computer Vision Project* (Roboflow, s.f).

i) Tomato Data Leaf 940 (TDL940)

El conjunto de datos integra un total de 2,255 de muestras digitales categorizadas bajo la clase hoja. La resolución espacial de cada imagen a color es de 640×640 píxeles. Los puntos por resaltar en este compendio son: I) presenta una alta variedad de dimensiones, morfología y especies (hoja del tipo sierra y papa), II) múltiples perspectivas ópticas, III) distintos niveles de oclusión, IV) transiciones cromáticas por el estado fitosanitario de la planta, V) fluctuaciones de iluminación natural, VI) presencia de diversos elementos del cultivo y VII) escenarios complejos en entornos de exteriores. Las imágenes se distribuyen en tres divisiones: entrenamiento (1,974), validación (187) y pruebas (94). En la Figura 4.9, se observan cuatro tomas.



Figura 4.9. Imágenes del conjunto *Tomato Data Leaf 940* (Roboflow, s.f).

j) Tomato Leaf Disease (TLD)

Esta recopilación contiene 4,132 imágenes de hojas de plantas de jitomate de varias especies en diferentes condiciones fitosanitarias. Los archivos se encuentran distribuidos en tres particiones independientes: entrenamiento (3,099), validación (867) y prueba (166) y clasificadas en 10 categorías de patologías.

Los atributos distintivos que destacar son los siguientes: I) extensa diversidad de alteraciones morfológicas (forma) y transiciones cromáticas (color) en la lámina foliar, II) presencia de oclusiones parciales y solapamientos entre las mismas estructuras de la planta, III) Múltiples perspectivas de captura que abarcan variaciones en los ángulos de incidencia y distancias respecto al objetivo, IV) exposición a diferentes condiciones de iluminación y V) escenas capturadas tanto en entornos controlados como en escenarios de campo abierto. La Figura 4.10 exhibe algunas tomas pertenecientes al banco de imágenes.



Figura 4.10. Imágenes del conjunto *TLD* (Roboflow, s.f).

4.2 Agrupación de imágenes web

El conjunto de imágenes está constituido por 1,562 ejemplares las cuales fueron redimensionadas a un tamaño de 416×416 píxeles. Las muestras provienen de varias fuentes web públicas como *blogs*, foros, redes sociales y repositorios de imágenes privados. La Figura 4.11, muestra ejemplos de la diversidad del *dataset* que, reúne algunas capturas hechas en escenarios complejos en exteriores.

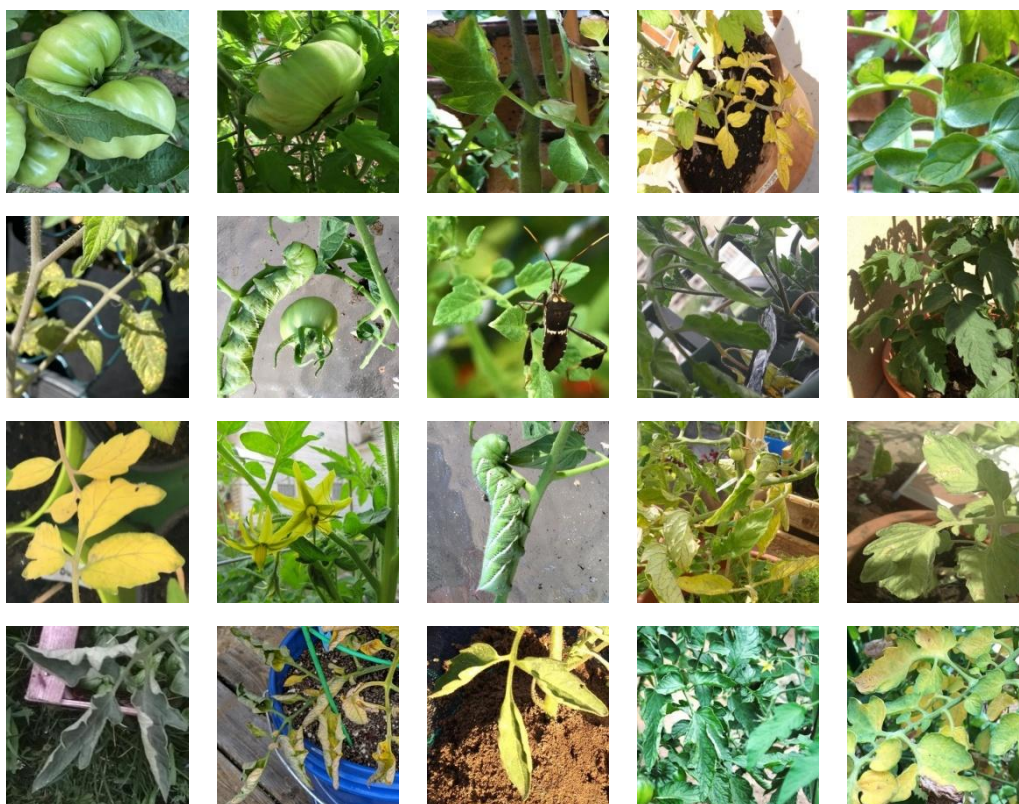


Figura 4.11. Collage de hojas de la planta del jitomate.

Las características distintivas de esta recolección son la presencia de elementos ajenos a las hojas de jitomate que no suelen estar presentes en los conjuntos de imágenes de *RoboFlow* tales como: gusanos, flores, macetas, varillas y otros elementos del cultivo. Asimismo, se incluyen imágenes con agrupaciones densas de hojas sanas y con anomalías que producen oclusiones, sombras y, una amplia variabilidad de iluminación, escalas y perspectivas de captura.

La incorporación de este material tiene como propósito incrementar la generalización del modelo al enriquecer el contenido del conjunto con escenarios complejos inherentes a las condiciones de campo abierto.

El etiquetado del conjunto propuesto se elaboró de manera manual usando la herramienta *Computer Vision Annotation Tool (CVAT, s.f)* en su versión gratuita. Respecto a las imágenes provenientes de *RoboFlow*, se llevó a cabo un proceso de revisión donde se corrigieron: I) las instancias mal definidas debido a una delimitación parcial o errónea de la hoja, II) agregar las láminas foliares omitidas en la anotación original y, III) la homologación del número de clases a una única categoría, debido a que el objetivo principal de esta investigación es la localización de regiones donde yacen hojas, tal y como se ilustra en la Figura 4.12.



Figura 4.12. Etiquetado de múltiples hojas.

La demarcación de cada lámina no contempla una etiqueta que identifique especie o estado de salud de la planta.

La Tabla 4.1 resume las especificaciones de las imágenes adquiridas a través de *Roboflow*, así como las del conjunto de datos propuesto.

Tabla 4.1. Composición del banco de imágenes propuesto.

Dataset	Núm. imágenes recolectadas	%	URL
<i>A4 [Roboflow, s.f]</i>	906	30.20%	http://bit.ly/4209xML
<i>ACVP [Roboflow, s.f]</i>	20	0.67%	https://bit.ly/4ikDUUU
<i>T7[Roboflow, s.f]</i>	44	1.47%	https://bit.ly/4bNEsAe
<i>TT [Roboflow, s.f]</i>	48	1.60%	https://bit.ly/4bD39zf
<i>DC [Roboflow, s.f]</i>	50	1.67%	https://bit.ly/4ij4UnR
<i>NCVP [Roboflow, s.f]</i>	64	2.13%	http://bit.ly/4bD3ulv
<i>TCVP [Roboflow, s.f]</i>	105	3.50%	https://bit.ly/4iAJ6nl
<i>TDL940 [Roboflow, s.f]*</i>	51	1.70%	https://universe.roboflow.com/mohsine-t8znx/tomatodataleaf940/dataset/1
<i>PCVP[Roboflow, s.f]</i>	100	3.33%	https://bit.ly/4iFOcig
<i>TDL[Roboflow, s.f]</i>	50	1.67%	https://bit.ly/4kALfRX
Agrupación de imágenes propuesto	1,562	52.06%	
	3,000 imágenes (27,397 instancias)	100%	*TDL940 ya no está disponible

Al contrastar el conjunto propuesto integrado por 3,000 imágenes con 27,397 instancias contra *PlantDoc* (Singh *et al.*, 2020) cuya colección es de 2,482 imágenes y 8,595 etiquetas, se observa una diferencia de 518 tomas digitales y 18,802 anotaciones; es decir, el presente trabajo incrementa sustancialmente la densidad de etiquetas.

Otro banco de imágenes de acceso libre es *FieldPlant* (Moupojou *et al.*, 2023) que contiene 5,156 ejemplares y 8,629 cajas delimitadoras. En este caso hay una mayor cantidad de imágenes respecto al conjunto de esta tesis; sin embargo, hay hojas que no están etiquetadas.

4.3 Conjunto para la etapa de pruebas

Se seleccionó un subconjunto de 13 imágenes representativas de las condiciones operacionales en cultivos a campo abierto. El conjunto de prueba abarca: I) diferentes condiciones de salud de la planta (aparentemente sanas y anómalas), II) diversidad en el entorno de fondo y III) variabilidad en escalas, perspectivas de captura y morfología foliar. Las dimensiones y los formatos de codificación (tales como *.jpg*, *.png* y *.webp*) de las imágenes originales se mantuvieron variables con el propósito de evaluar la capacidad de generalización del modelo durante la fase de inferencia. En la Figura 4.12 se ilustran los 13 ejemplares.



Figura 4.13. Collage del conjunto de pruebas.

4.4 Conjunto de imágenes de otros cultivos

Con el objetivo de evaluar la generalización del modelo para localizar las hojas presentes en una imagen, se consideraron otros cultivos. El contenido en el dataset CWD30 (Ilyas *et al*, 2025) se conforma de 219,770 imágenes que se organiza en dos categorías: I) 10 cultivos importantes en la producción a nivel global y regional de Corea y, II) 20 especies de malezas más problemáticas del mundo, de acuerdo con los autores.

Los cultivos seleccionados para su estudio son cacahuete, frijol y frijol rojo que presentan hojas de forma ovalada a romboidal (Esaú, 1977). Otros cultivos se descartaron por presentar hojas del tipo lanceoladas; es decir, alargadas y estrechas como las hojas del maíz y mijo (Valla, 2007). En la Figura 4.13 se muestran tres cultivos capturados en zonas de cultivo ubicadas al exterior.



Cacahuete **Frijol** **Frijol rojo**
Figura 4.14. Ejemplo de cultivos seleccionados (Ilyas *et al*, 2025).

Se seleccionó el cultivo de frijol porque en base en la fuente (SAGARPA, 2017), el frijol representa un cultivo clave en la dieta de los mexicanos con un consumo per cápita anual de 9.9 Kg. México ocupa el cuarto lugar a nivel mundial en la producción de esta leguminosa con el 5.50%. Por otro lado, de acuerdo con (SAGARPA, 2023), el valor de producción del cacahuate en México fue de un millón 304 mil 192 miles de pesos a nivel nacional, de los cuales 376 mil 766 miles de pesos fueron de Sinaloa (principal productor). El estado de Morelos figura entre los 10 principales productores del cacahuate (INIFAP, 2024).

Cada categoría del conjunto de imágenes se subclasifica acorde al número de semana en desarrollo desde la semana uno a la ocho. Los ejemplares tienen una resolución en promedio de $4,032 \times 3,024$ píxeles que son capturados directamente en los campos y en áreas específicas de cultivo a la intemperie, donde ocurren variaciones de iluminación debido a que la adquisición se realizó en las horas diurnas; también presentan variaciones en las condiciones climáticas y estacionales. Los ángulos de visión con respecto al objetivo son de 0° , 45° y 90° . *CWD30* no cuenta con un *GT* orientado para tareas de detección o segmentación de láminas foliares.

En las Tablas 4.2, 4.3 y 4.4, se detalla la distribución de las imágenes consideradas en este estudio tomando en cuenta la semana de vida del cultivo y el ángulo de captura del cacahuate, frijol y frijol rojo.

Tabla 4.2. Distribución de ejemplares del cacahuate.

Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
1	366	880	322
2	78	905	409
3	74	387	499
4	0	779	394
5	0	556	312
6	0	221	150
7	0	174	142
8	0	48	47

Tabla 4.3. Distribución de ejemplares del frijol.

Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
1	705	783	299
2	282	776	384
3	96	570	285
4	0	841	396
5	0	284	136
6	0	66	89
7	0	58	66
8	0	23	25

Tabla 4.4. Distribución de ejemplares del frijol rojo.

Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
1	630	706	371
2	217	1,100	449
3	48	540	275
4	96	570	284
5	0	414	274
6	0	106	77
7	0	31	48
8	0	40	49

CAPÍTULO 5

MÉTODO DE SOLUCIÓN

En el capítulo se detallan las modificaciones realizadas en el modelo que realiza la localización y su unión con la etapa de segmentación de hojas de la planta del jitomate en imágenes capturadas en entornos exteriores. En la fase de localización se encuentra la principal contribución de esta tesis al realizar la modificación a la red *Yolov11n* con el propósito de reducir los errores de inferencia del tipo falso negativo y mejorar la precisión y ajuste de las cajas delimitadoras.

5.1 Etapas del método propuesto

El flujo de trabajo planteado tiene como objetivo automatizar la localización y segmentación de las múltiples estructuras foliares presentes en imágenes capturadas en condiciones de campo abierto. Esta estrategia se estructura en dos etapas secuenciales: localización y segmentación (ver Figura 5.1). Para su ejecución, se emplean como base dos redes neuronales de aprendizaje profundo especializadas, donde cada red resuelve una tarea específica del proceso.

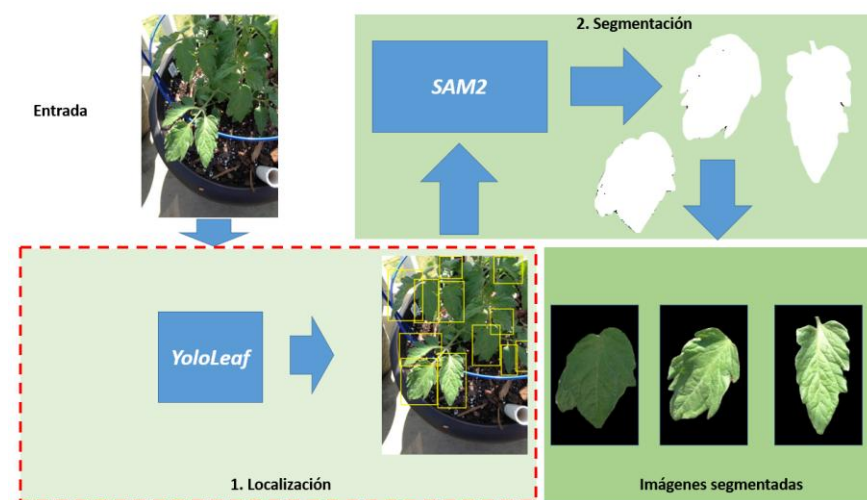


Figura 5.1. Flujo de trabajo de la propuesta de solución.

La fase de localización se centra en maximizar la tasa de detección de las hojas presentes en la imagen, determinando la posición de cada una de ellas mediante la generación de cajas delimitadoras (*Bounding Boxes*) que acotan la región de interés.

En esta etapa se introduce una estrategia orientada al refinamiento de los mapas de características (*feature maps*) y a la optimización de la localización de hojas afectadas por factores bióticos y abióticos. Dicha propuesta toma como base la arquitectura de la red

Yolov11n, implementando modificaciones estructurales tanto en el componente de extracción de características (*backbone*) como en la función de pérdida por regresión original. Los elementos clave se esquematizan en la Figura 5.2.

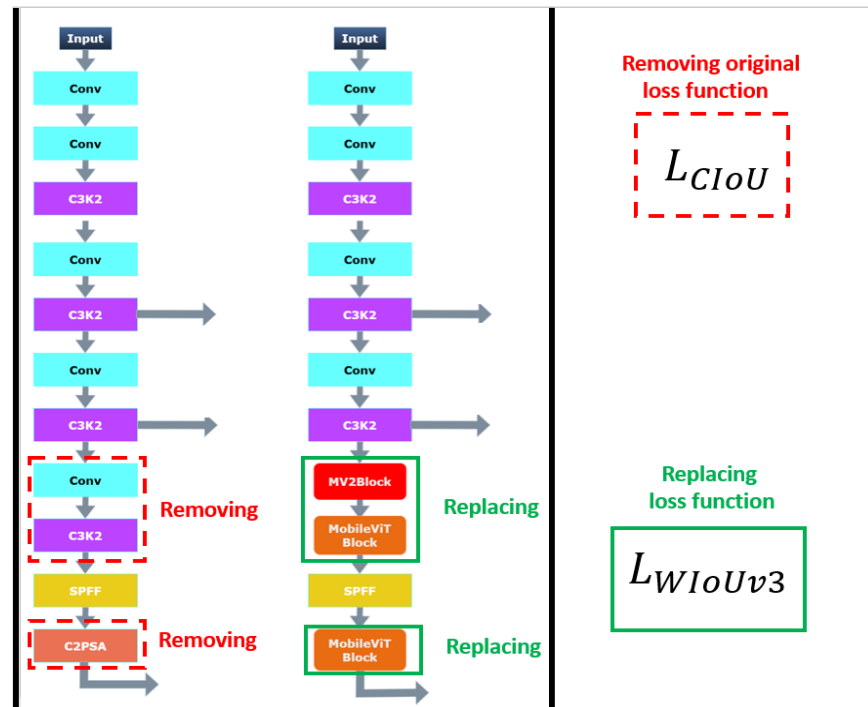


Figura 5.2. *Backbone* original y modificado; remplazo de la función de regresión.

Las modificaciones estructurales se localizan específicamente en el *backbone* de la red, con el objetivo de optimizar la extracción de características y, simultáneamente, mitigar la carga computacional medida en *FLOPs* (operaciones de punto flotante). Para ello se removieron los módulos originales: *Conv*, *C3k2* y *C2PSA* ubicados en los índices siete, ocho y diez del *backbone*, respectivamente. En su lugar, los bloques *MV2Block* y *MobileViTBlock* son implementados.

El módulo *MV2Block* sustituye a la capa *Conv* con el fin de reducir el coste computacional, mientras que componente *MobileViTBlock* reemplaza a *C3k2* y *C2PSA* para refinar los mapas de características mediante la fusión de características locales y globales calculadas por las operaciones convolucionales y por los transformadores de visión respectivamente. Además, la integración de este bloque en dicha posición ayuda a reducir la pérdida de resolución espacial en el módulo *SPPF* conformado por una agrupación de capas *Pooling*. (Parez et al., 2023) señalan que la reducción dimensional a través de las capas *Pooling* en los modelos *CNN* provoca la pérdida de información crítica, comprometiendo la localización precisa de las características más prominentes.

La sustitución de la función de pérdida original *CioU* por la variante *WIoUv3* (Tong et al., 2023) tiene como finalidad optimizar la localización de las estructuras foliares de interés (hojas). *WIoUv3* evalúa la calidad de las cajas delimitadoras predichas o *Anchor Boxes* (ABs) e implementa un mecanismo de enfoque adaptativo para concentrar los esfuerzos de optimización estratégicamente en aquellas muestras de calidad intermedia. Esta estrategia

guía al modelo hacia una delimitación geométrica más precisa de la región ocupada por las hojas, mitigando el impacto de predicciones ruidosas durante el entrenamiento.

Finalmente, el módulo de segmentación incorpora la arquitectura ligera *SAM2-Tiny* (Ravi et al., 2024), la cual recibe las coordenadas de los cuadros delimitadores inferidas por *YoloLeaf* (el modelo propuesto) en forma de *prompts* de entrada geométricos. Dichos indicadores acotan las regiones de interés (*ROI*) donde se ubican las estructuras foliares, permitiendo que el modelo generativo ejecute la extracción de la máscara binaria a nivel de píxel correspondiente a cada hoja.

5.2 Modificaciones en la arquitectura de la red *Yolo11n*

En las Tablas 5.1 y 5.2 se sintetizan los componentes constituyentes en *Yolov11n* y *MobileViT* respectivamente.

Tabla 5.1. Componentes originales en *Yolo11n*.

Componente	Descripción
<i>Conv</i>	Generación de mapas de características.
<i>C3k2</i>	Refinación de los mapas de características.
<i>C2PSA</i>	Análisis de regiones clave en los mapas de características.

Tabla 5.2. Componentes originales en *MobileViT*.

Componente	Descripción
<i>MV2Block</i>	Creación de mapas de características.
<i>MobileViTBlock</i>	Fusión de características locales y globales.

En la Tabla 5.3 se detallan los cambios realizados en el *backbone* de la red base *Yolov11n*, así como las modificaciones efectuadas en *MV2Block* y *MobileViTBlock*.

Tabla 5.3. Configuraciones en *Yolo11n*.

Componente original	Justificación
<i>Conv</i>	Se sustituye por el bloque <i>MV2Block</i> para definir nuevos mapas de características a un menor costo computacional.
<i>C3k2</i> y <i>C2PSA</i>	Se cambia por el módulo <i>MobileViTBlock</i> para afinar los mapas mediante la fusión estratégica de relaciones locales y globales.
<i>MV2Block</i>	Se anexa <i>Convolution Block Attention Module</i> (<i>CBAM</i>) para mejorar la representación de características en los mapas generados.
<i>MobileViTBlock</i>	Se eliminan las capas <i>Conv</i> (convoluciones estándar) para reducir el costo computacional.

Con el propósito de facilitar la interpretación de los módulos propuestos, la Tabla 5.4, se introduce una nomenclatura visual por colores y la caracterización funcional de los componentes que integran cada bloque de la red.

Tabla 5.4. Código de colores.

Color	Nombre	Descripción
	<i>Point Wise Convolution (PWC)</i>	Efectúa la fusión de mapas de características empleando un <i>kernel</i> $k = 1$.
	<i>Standar Convolution (SC)</i>	Realiza la creación de mapas de características usando un <i>kernel</i> $k = 3$.
	<i>CBAM</i>	Enfatiza y suprime información relevante y no significativa dado un conjunto de mapas de características.
	<i>Transformer</i>	Calcula las relaciones globales existentes a partir de un conjunto de mapas de características.
	<i>BatchNorm</i>	Reduce la escala numérica de los píxeles que componen un mapa de características para un cálculo eficiente.
	<i>SiLu</i>	Función de activación no monótona implementada para eliminar patrones lineales en la información durante el entrenamiento.

a) *MV2Block*

El propósito de este módulo es refinar y optimizar los mapas de características con un menor costo computacional, aprovechando que las convoluciones separables en profundidad (*Depthwise Separable Convolutions*, DWC) reducen significativamente los *GFLOPs*. Asimismo, se incorpora el mecanismo de atención *CBAM* (*Convolutional Block Attention Module*) (Woo et al., 2018), diseñado para potenciar la representatividad y la inferencia en arquitecturas convolucionales enfocadas en la detección de objetos. En la Figura 5.3 se ilustra esta transición: el inciso a) presenta de forma simplificada la configuración interna del bloque *MV2Block* original, mientras que el inciso b) se presenta la modificación propuesta en esta investigación.

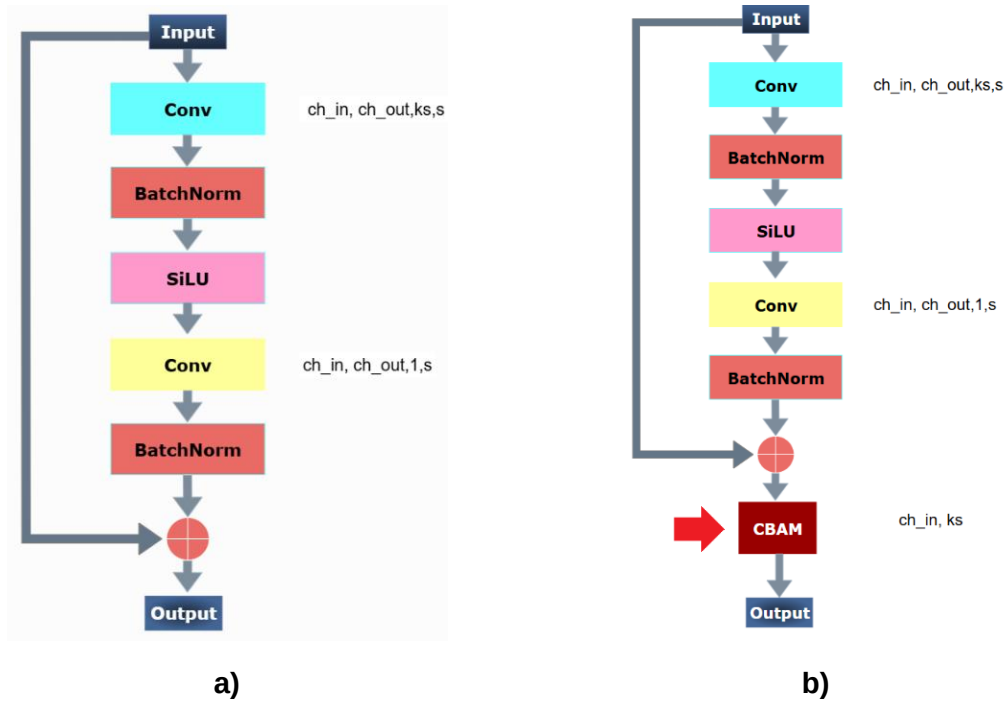


Figura 5.3. La figura esquematiza: a) módulo original *MV2Block* y b) inclusión de *CBAM* en *MV2Block*.

El bloque *MV2Block* modificado recibe un tensor de entrada $X \in \mathbb{R}^{H \times W \times C}$ que aloja información local proveniente del módulo *C3K2* propia de la arquitectura *Yolov11n*. El tensor X es procesado por la capa de convolución estándar *Conv* (SC) para la generación de los mapas de características preliminares $X_{conv} \in \mathbb{R}^{H \times W \times C}$. Simultáneamente, los valores de X_{conv} se normalizan mediante una capa de *Batch Normalization* (BN) con el fin de estabilizar la distribución de las activaciones y optimizar la convergencia durante la introducción de no linealidades en el mapa de salida $X_{convN} \in \mathbb{R}^{H \times W \times C}$. Posteriormente, la capa de convolución separable en profundidad *Conv* (DWC) ejecuta la transformación espacial de los mapas normalizados, reduciendo la dimensión del nuevo mapa de características resultante $X_{DWC} \in \mathbb{R}^{H \times W \times C}$.

Posteriormente, el tensor de entrada original X y el mapa de características procesado X_{DWC} se adicionan mediante una conexión residual ($X'_{DWC} \in \mathbb{R}^{H \times W \times C}$) antes de ser procesados por el módulo *CBAM*. El tensor combinado es transferido al módulo *CBAM*, el cual aplica un refinamiento adaptativo en el flujo de datos. Dicho proceso consiste en explorar las relaciones inter-canal como las dependencias espaciales para amplificar las características geométricas de la hoja y atenuar la información redundante, generando finalmente el mapa refinado $Y_{CBAM} \in \mathbb{R}^{H \times W \times C}$.

b) *MobileViTBlock*

Las modificaciones introducidas en el bloque *MobileViTBlock* tiene como finalidad reducir el número de parámetros de entrenamiento y generar mapas refinados a partir de la convergencia de información local y global, calculada mediante operaciones convolucionales y mecanismos de autoatención.

Para ello, se elimina la primera capa SC y el módulo final se sustituyó por una convolución en punto (*PWC*). Al omitir la redundancia en la generación de mapas intermedios y optimizar la agregación de los descriptores espaciales previamente calculados. Asimismo, la conexión residual preserva el flujo de información que proviene de las salidas del módulo *MV2Block* (índice ocho) y *SPFF* (índice diez) concatenándose con la salida del *Transformer* para consolidar una eficiente fusión multiescala.

En la Figura 5.4 se presentan: a) la configuración interna del bloque *MobileViTBlock* original y b) los cambios propuestos para dicho módulo.

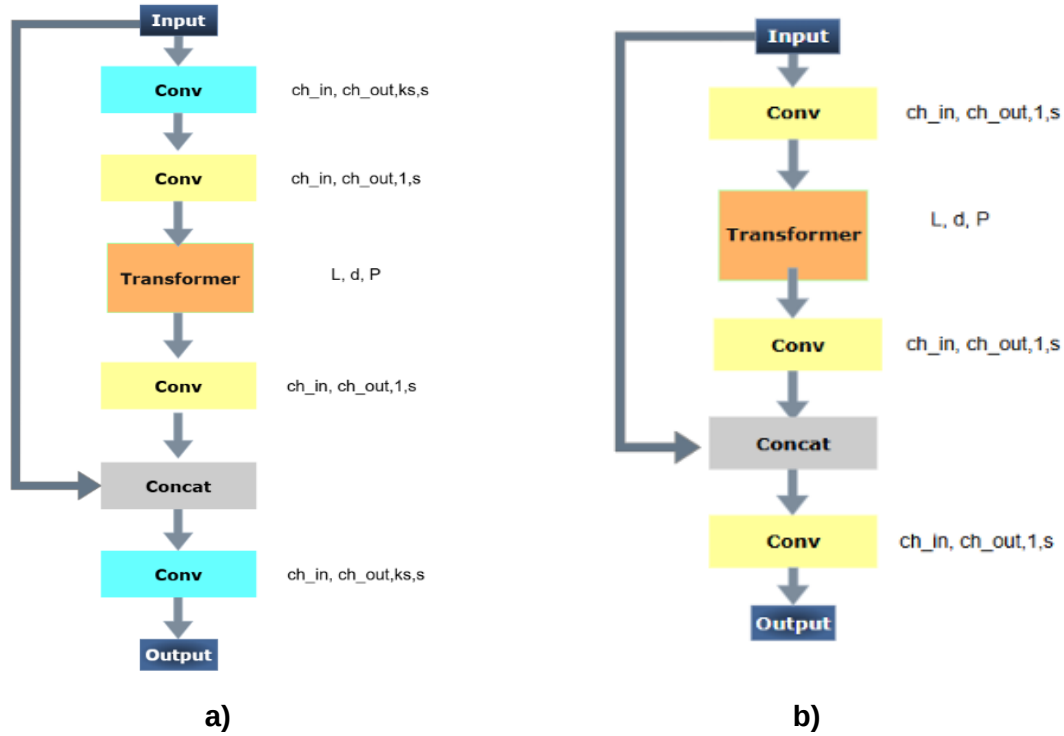


Figura 5.4. La figura ilustra: a) módulo original *MobileViTBlock* y b) remoción de capas *Conv* (SC) en el módulo *MobileViTBlock*.

El bloque *MobileViTBlock* modificado recibe un tensor de entrada $X \in \mathbb{R}^{H \times W \times C}$ que contiene la información espacial local proveniente del bloque *MV2Block* o *SPFF*. X es procesado por la capa *Conv* (*PWC*) para producir $X_L \in \mathbb{R}^{H \times W \times d}$ que aloja las proyecciones dimensionales espaciales d (donde $d > C$) a partir de la combinación de los canales de entrada, para el modelado de dependencias no locales de largo alcance a través del bloque transformador (*Transformer*).

Para el cálculo de las representaciones globales, el tensor X_L se procesa para obtener N *non-overlapping flattened patches* $X_U \in \mathbb{R}^{P \times N \times d}$, donde $P = wh$, $N = \frac{HW}{P}$ determina el número de parches; $h \leq n$ y $w \leq n$ corresponden al alto y ancho de cada parche respectivo. Las relaciones espaciales de cada parche $p \in \{1, \dots, P\}$, son codificados al emplear transformadores para obtener $X_G \in \mathbb{R}^{P \times N \times d}$:

$$X_G(p) = \text{Transformer}(X_U(p)), 1 \leq p \leq P$$

Dentro del transformador se define un conjunto de codificadores l que a su vez se subdividen en $MHSA$ y MLP . A partir de las proyecciones lineales de $X_U(p)$, se efectúa el cálculo de la atención (Ecuación 5.1) que consiste en calcular el producto punto de $Query$ Q con respecto a $Keys$ K , sobre la raíz cuadrada de d , evaluado por la función $SoftMax$ y multiplicado por $Value$ V .

$$Atencion(Q, K, V) = SoftMax\left(\frac{QK^T}{\sqrt{d}}\right) V \quad \text{Ecuación 5.1}$$

Donde las matrices W^Q , W^K , y W^V alojan los parámetros que se entrenan y determinan el peso asignado a $Query$, Key , y $Value$ respectivamente.

$X_G \in \mathbb{R}^{P \times N \times d}$ se procesa hasta obtener $X_F \in \mathbb{R}^{H \times W \times C}$ que proyecta al espacio de dimensión C utilizando la capa PWC y, posteriormente se produce $X_C = Concat(X_F, X) \in \mathbb{R}^{H \times W \times 2C}$.

Finalmente, la última capa PWC se emplea para fusionar las características locales y globales presentes en X_C respectivamente, generando el mapa de características enriquecido $X_R \in \mathbb{R}^{H \times W \times C}$.

En la Tabla 5.5 se describen las características principales de los nuevos componentes en el *backbone* propuesto como: el módulo, ubicación en la red (capa), número de parámetros, *MegaFlops* (*MFlops*), argumentos y valores definidos. La colocación de los módulos se aborda en el capítulo siete (estudio de ablaciones). La Tabla 5.6 detalla los módulos originales del *backbone* de *Yolov11n*.

Tabla 5.5. Módulos del *backbone* en la arquitectura propuesta.

Módulo	Capa	Parámetros	MFlops	Argumentos	Valores
MV2Block	7 (<i>Backbone</i>)	34,658	3.10	<i>ch_in</i> : input channels <i>ch_out</i> : output channels <i>ks</i> : kernel size <i>s</i> : steps <i>e</i> : expansion	<i>ch_in</i> : 128 <i>ch_out</i> : 128 <i>ks</i> : 7 <i>s</i> : 2 <i>e</i> : 1
MobileViTBlock	8 (<i>Backbone</i>)	83,648	17.85	<i>ch_in</i> : input channels <i>ch_out</i> : output channels <i>ks</i> : kernel size <i>L</i> : concatenation <i>d</i> : dimension of vector <i>P</i> : size of patch	<i>ch_in</i> : 128 <i>ch_out</i> : 128 <i>ks</i> : 7 <i>L</i> : 3 <i>d</i> : 96 <i>P</i> : 1
MobileViTBlock	10 (<i>Backbone</i>)	190,656	35.94	<i>ch_in</i> : input channels <i>ch_out</i> : output channels <i>ks</i> : kernel size <i>L</i> : concatenation <i>d</i> : dimension of vector <i>P</i> : size of patch	<i>ch_in</i> : 256 <i>ch_out</i> : 256 <i>ks</i> : 3 <i>L</i> : 3 <i>d</i> : 96 <i>P</i> : 1
		Total: 308,962	Total: 56.89		

Tabla 5.6. Módulos originales del *backbone* en la arquitectura *Yolov11n*.

Módulo	Capa	Parámetros	MFlops	Argumentos	Valores
Conv	7 (Backbone)	295,424	49.92	<i>ch_in</i> : input channels <i>ch_out</i> : output channels <i>ks</i> : kernel size <i>s</i> : steps	<i>ch_in</i> : 128 <i>ch_out</i> : 256 <i>ks</i> : 3 <i>s</i> : 2
C3k2	8 (Backbone)	346,112	58.49	<i>ch_in</i> : input channels <i>ch_out</i> : output channels <i>n</i> : number of blocks	<i>ch_in</i> : 256 <i>ch_out</i> : 256 <i>n</i> : 1 <i>shortcut</i> : True
C2PSA	10 (Backbone)	249,728	47.68	<i>ch_in</i> : input channels <i>ch_out</i> : output channels	<i>ch_in</i> : 256 <i>ch_out</i> : 256
		Total: 891,264	Total: 156.09		

De acuerdo con el análisis, la modificación en la arquitectura beneficia a la disminución tanto en el número de parámetros de entrenamiento como en las operaciones de punto flotante.

5.3 Sustitución de la función *CIoU* de la red *Yolo11n*

La función que estima la ubicación de las cajas delimitadoras por defecto en la red *Yolov11n* es *CIoU* (*Complete Intersection over Union*) (Ecuación 5.2); en la Figura 5.5, se ilustran las variables que participan en *CIoU*. Donde h y w corresponden al alto y ancho de *Predicted Box*; h^{gt} y w^{gt} son el alto y ancho de *Real Box*; c_h y c_w se refieren al alto y ancho del mínimo cerramiento formado a partir de *Predicted Box* y *Real Box*; $b_{c_x}^{gt}$, $b_{c_y}^{gt}$ y b_{c_x} , b_{c_y} indican el punto central del cuadro delimitador del *ground truth* y el predicho respectivamente; (se describen las variables faltantes en la ecuación 5.2)

$$L_{CIoU} = 1 - IoU + \frac{(b_{c_x}^{gt} - b_{c_x})^2 + (b_{c_y}^{gt} - b_{c_y})^2}{(c_w)^2 + (c_h)^2} + \frac{4}{\pi^2} \left(\tan^{-1} \frac{w^{gt}}{h^{gt}} - \tan^{-1} \frac{w}{h} \right) \quad \text{Ecuación 5.2}$$

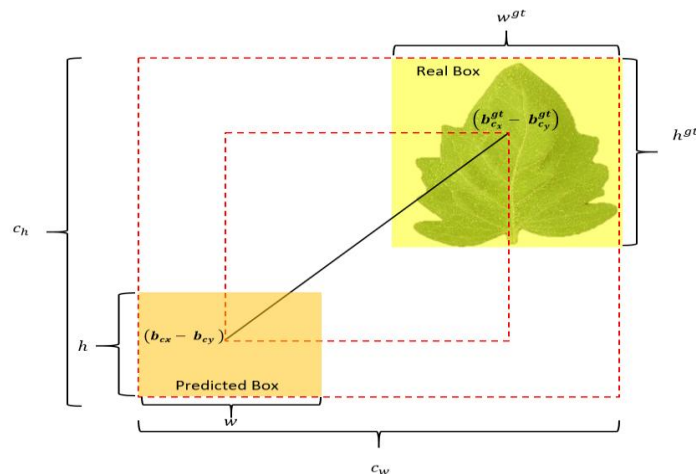


Figura 5.5. Representación gráfica de las variables.

La función original *CIoU* (*Complete Intersection over Union*) integrada en *Yolov11n* presenta los siguientes inconvenientes: a) no se considera el número de traslapes entre los *BBs* que se generan, b) si los valores del ancho y alto son diferentes, la función de penalización calcula la diferencia real entre el dato real e inferido, pero cuando la relación de la forma del cuadro real y el cuadro predicho son iguales y, c) el cálculo de *CIoU* implica una función trigonométrica lo que significa, un aumento en el costo computacional (Wang *et al.*, 2023).

La Figura 5.6 ejemplifica un escenario de un entorno al exterior, donde se observan múltiples tamaños, formas y traslapes de los cuadros delimitadores sobre las hojas de la planta del jitomate.



Figura 5.6. Aspectos y solapamientos de instancias delimitadoras.

Ante las limitantes que tiene la función *CIoU* debido a la naturaleza del problema bajo estudio, es necesario replantear el cálculo de las coordenadas de los *BBs* para reducir las imprecisiones al localizar los objetivos de interés presentes en la imagen. Esta reconsideración se hace al implementar la función *WIoUv3* en lugar de *CIoU* para una mejor definición de los cuadros delimitadores.

WIoUv3 al basarse en *WIoUv1* se preserva la propiedad de emplear la distancia de centroides que actúa como mecanismo de atención de modo que cuando ambos cuadros delimitadores se traslapan la penalización disminuye, contribuyendo a una mejor inferencia. Además, *WIoUv3* se enfoca en discriminar aquellos *ABs* atípicos que dan origen a la instancia delimitadora principal (*BB*). En la Figura 5.7 se representan los elementos que definen la región de interés donde figura una hoja.

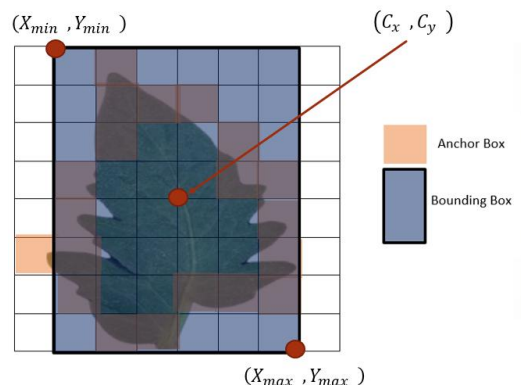


Figura 5.7. Esquematación gráfica del *AB* y *BB*.

Matemáticamente, la instancia delimitadora $BB = (X_{min}, Y_{min}, X_{max}, Y_{max})$, existe en la imagen I sujeta a $0 \leq X_{min} < X_{max} \leq W$; $0 \leq Y_{min} < Y_{max} \leq H$; donde W y H son el ancho y alto de I . Un *Anchor Box* se define como $AB_{i,j,k} = (C_x^{i,j}, C_y^{i,j}, w_k, h_k)$, $k = 1 \dots K$ donde $C_x = \frac{X_{min} + X_{max}}{2}$; $C_y = \frac{Y_{min} + Y_{max}}{2}$; $w = X_{max} - X_{min}$; $h = Y_{max} - Y_{min}$; w y h son el alto y ancho de AB_k , que es la representación mínima de una región en I .

WIoUv3 al basarse la función focal monótona propuesta en *WIoUv2*, se introduce el uso de los hiperparámetros α y δ que definen el comportamiento no monótono de la nueva función. En la Figura 5.8 se ejemplifican las tres funciones estudiadas en esta investigación.

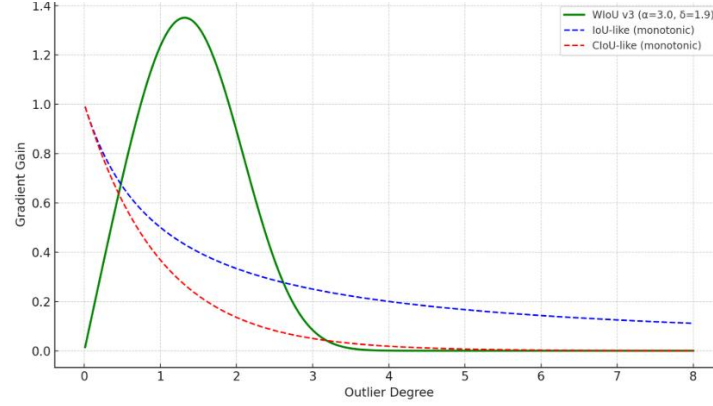


Figura 5.8. Comparación del comportamiento monótono y no monótono.

Nótese como el par de funciones *CIoU* e *IoU* de izquierda a derecha decrece mientras que, *WIoUv3* presenta cambios positivos y negativos en el valor de la pendiente. Este comportamiento no monótono define la estrategia de ganancia del gradiente que optimiza dinámicamente la calidad de AB_k .

Cuando la pendiente es negativa (muestras muy malas o *outliers*), la reducción de las ganancias r evita que los ejemplos atípicos β perjudiquen el entrenamiento. Cuando la pendiente es positiva (muestras muy buenas), se asignan ganancias pequeñas r para evitar un sobreajuste. Por lo tanto, el enfoque neutral se centra en las muestras de "calidad ordinaria" (cerca del pico de la curva), que son las que realmente ayudan a mejorar el rendimiento de localización de objetivos en la imagen.

Los AB_k que pertenecen al intervalo $[4,8]$ no afectan negativamente el gradiente, contribuyendo a que el modelo se enfoque en la calidad promedio de los AB_k en $[0,3]$.

Los valores numéricos de $\alpha = 3.0$ y $\delta = 1.9$ son tomados como referencia de (Wang G. et al., 2023), donde se reporta una mejora en las métricas de: *Recall*, *mAP@50* y *mAP@50-95* utilizando en *Yolov8s*. En el capítulo siete, estudio de ablaciones se confirma una mejora en los indicadores *Recall* y *F1-Score* (ver Tabla 7.4).

CAPÍTULO 6

ANÁLISIS Y DISEÑO DEL SISTEMA

En el capítulo se presentan los diagramas funcionales y el diseño de la interfaz de la aplicación móvil desarrollada para la localización y segmentación de hojas de plantas de jitomates en imágenes. Dicho desarrollo fundamenta su operación en el despliegue conjunto de los modelos *YoloLeaf* y *MobileSAM*, especializados en los dominios de detección y segmentación.

6.1 Requerimientos funcionales

A continuación, se enlistan y se describen los cuatro requerimientos funcionales de la aplicación móvil:

- **Lectura archivos:** acceso a los directorios del dispositivo móvil para la selección del archivo a procesar (imagen).
- **Proceso de localización:** estimación de la ubicación y extensión espacial de cada objeto presente en el ejemplar.
- **Proceso de segmentación:** inferencia de la máscara de píxeles de cada objetivo previamente localizado y acotado por el cuadro delimitador.
- **Escritura de archivos:** ingreso al fichero destino para el almacenamiento de las imágenes segmentadas en extensión *png*.

6.2 Casos de uso

La Figura 6.1 ilustra las tres maneras en que el usuario interactúa una vez ejecutada la aplicación móvil.

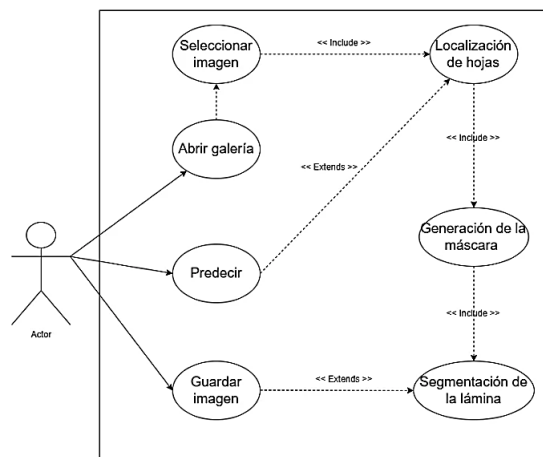


Figura 6.1. Diagrama de casos de uso.

Al iniciar la aplicación, el usuario accede a la galería e importa la imagen a procesar desde la galería del dispositivo. Posteriormente, el modelo *YoloLeaf* localiza las hojas mediante la predicción de cajas delimitadoras (*Bounding Boxes*). A continuación, se toma la información espacial de la posición de dicha caja como entrada para efectuar la delimitación precisa de la estructura foliar.

El usuario inspecciona la calidad de la imagen segmentada cualitativamente, decidir si la muestra es la correcta y, posteriormente guardar la nueva imagen segmentada en el dispositivo.

a) Diagramas de flujo

En las Figuras 6.2 y 6.3, se presentan los diagramas de flujo del procedimiento de localización y segmentación, respectivamente, que realizan dichos modelos. Las Tablas 6.1 y 6.2 especifican los procesos que conforman cada diagrama de flujo respectivo.

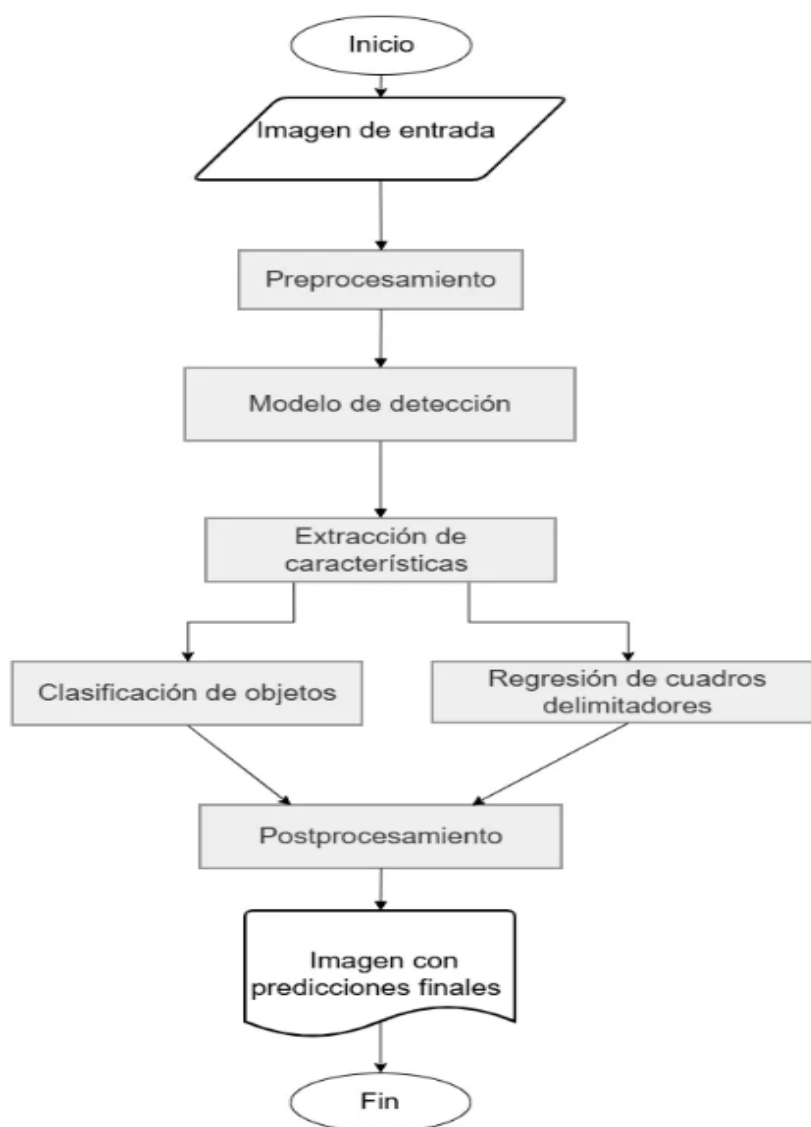


Figura 6.2. Diagrama de flujo del proceso del proceso de localización.

Tabla 6.1. Explicación de cada etapa del proceso de detección.

Proceso	Descripción
Imagen de entrada	Se recibe una captura digital donde figuran láminas foliares de la planta del jitomate obtenida a campo abierto.
Preprocesamiento	Para estandarizar la entrada a la red, se realiza un redimensionamiento espacial de la imagen y la normalización de los valores de intensidad de los píxeles.
Modelo de detección	A partir de la imagen normalizada, <i>YoloLeaf</i> localiza las hojas de jitomate presentes en la escena analizada.
Extracción de características	El modelo crea múltiples mapas de características mediante las capas convolucionales y bloques de atención. Estos últimos se encargan de enfatizar las características discriminativas y suprimir los componentes irrelevantes, lo que genera una representación enriquecida de los datos que conduce a una mejor inferencia.
Clasificación de objetos	A partir del mapa de características refinado se estima la probabilidad de pertenencia del objeto a las clases hoja o fondo.
Regresión de cuadros delimitadores	Se calculan las proporciones dimensionales de las cajas contenedoras de la región de interés asociada a cada hoja.
Postprocesamiento	Las dimensiones de las instancias localizadas se reescalan y proyectan para adaptarse a la resolución espacial de la imagen original de entrada.
Imagen con predicciones finales	Las predicciones del modelo se superponen sobre la imagen original, permitiendo la visualización de los cuadros delimitadores que localizan las múltiples hojas existentes.

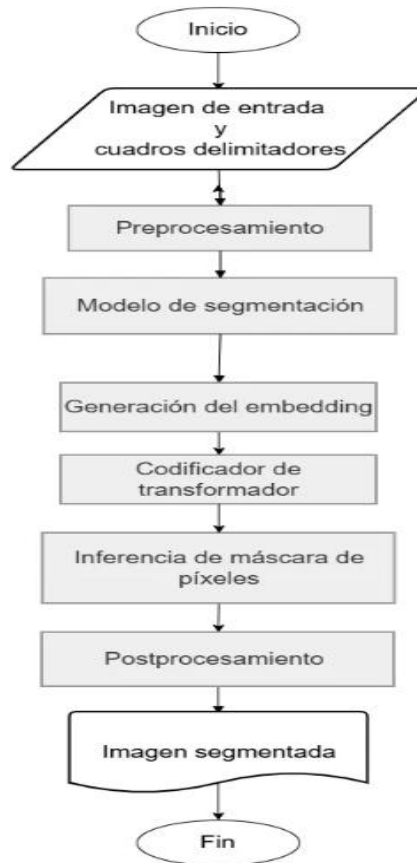


Figura 6.3. Diagrama de flujo del proceso de segmentación.

Tabla 6.2. Explicación de cada etapa del proceso de segmentación.

Proceso	Descripción
Imagen de entrada y cuadros delimitadores	Se recibe la imagen en la cual las láminas foliares de la planta del jitomate se encuentran espacialmente delimitadas.
Preprocesamiento	Se realiza el redimensionamiento geométrico de la imagen de entrada y, la normalización de las coordenadas correspondientes de la caja delimitadora.
Modelo de segmentación	<i>MobileSAM</i> emplea las dimensiones espaciales del <i>Bounding Box</i> como <i>prompt</i> de entrada que define la región de interés a procesar.
Generación del <i>embedding</i>	Creación de una representación de la información dentro de un espacio vectorial.
Codificador de transformador	Se efectúa el producto punto entre vectores para obtener la matriz de atención.
Predicción de la máscara de píxeles	Se gráfica una región que se asemeja a la forma del objeto de interés (hoja).
Postprocesamiento	La zona definida previamente se escala con respecto al tamaño original de la imagen, se preserva únicamente la región de interés con respecto al área inferida.
Imagen segmentada	El nuevo ejemplar ilustra una lámina foliar de la planta del jitomate sobre un fondo de color negro.

6.2 Maquetado del sistema móvil

La Figura 6.4 muestra una vista parcial de la imagen resultante y sus componentes para su uso y operaciones.



Figura 6.4. Maquetado de la aplicación móvil.

1. **Visualizador 1:** Muestra la imagen seleccionada y un ejemplo de una hoja detectada por medio de la caja delimitadora color rojo.
2. **Indicador:** Contabiliza el total de número de hojas detectadas.
3. **Visualizador 3:** Ilustra la lámina foliar delimitada previamente localizada.
4. **Indicador:** Señala el índice numérico de hoja segmentada.
5. **Botón Cargar:** Abre la galería de imágenes para explorar y seleccionar un ejemplar.
6. **Botón Predecir:** Infiere una nueva localización y segmentación de la hoja.
7. **Botón Guardar:** Almacena la hoja delimitada que figura en el visualizador 3.

CAPÍTULO 7

EXPERIMENTACIÓN Y RESULTADOS

Este capítulo detalla la configuración implementada en los diversos escenarios de la fase de experimental, especificando los valores de los hiperparámetros y la distribución de las particiones del conjunto de datos. Adicionalmente, se presenta el estudio de ablación, el proceso de validación cruzada y el análisis de las métricas de evaluación orientadas a la localización y segmentación de las láminas foliares.

7.1 Experimentación

Los experimentos se dividen en dos etapas:

- **Detección espacial:** El objetivo de la primera etapa se centra maximizar la tasa de localización de estructuras foliares a partir de imágenes adquiridas en condiciones de campo a cielo abierto. Para ello, se implementan y comparan los modelos *Yolov11n*, *Yolov12n* y la variante propuesta *YoloLeaf*, llevando a cabo las respectivas evaluaciones cuantitativas y cualitativas del rendimiento.
- **Segmentación de instancias:** El propósito de esta fase es extraer el contorno exacto de las láminas foliares localizadas por *YoloLeaf* en la etapa previa. Para este fin, se examina el comportamiento de las redes neuronales *SAM2 Tiny* y *MobileSAM*, diseñadas específicamente para la segmentación eficiente de imágenes.

Con respecto a la primera fase, y con el propósito de garantizar condiciones de paridad experimental que permitan validar la mejora propuesta por *YoloLeaf* ante el problema de la detección foliar, se detallan a continuación el conjunto de datos visuales y la configuración de hiperparámetros empleada de manera uniforme en las tres arquitecturas bajo análisis.

a) Banco de imágenes

La Tabla 7.1 detalla la distribución de 3,000 muestras digitales y 27,397 instancias existentes en los conjuntos de entrenamiento y validación establecidos. Se decide establecer una partición de datos 80/20 ya que, es la más empleada con base en los trabajos estudiados en el capítulo 2 (ver Tabla 2.1).

Tabla 7.1. Distribución de los ejemplares.

Entrenamiento (80%)	Validación (20%)
2,400 (imágenes)	600 (imágenes)
(22,013 instancias)	(5,384 instancias)
Total: 3,000 imágenes (27,397 instancias)	

La Figura 7.1, presenta una serie de gráficos que permiten analizar las propiedades geométricas y la distribución estadística de las cajas delimitadoras del banco de imágenes propuesto.

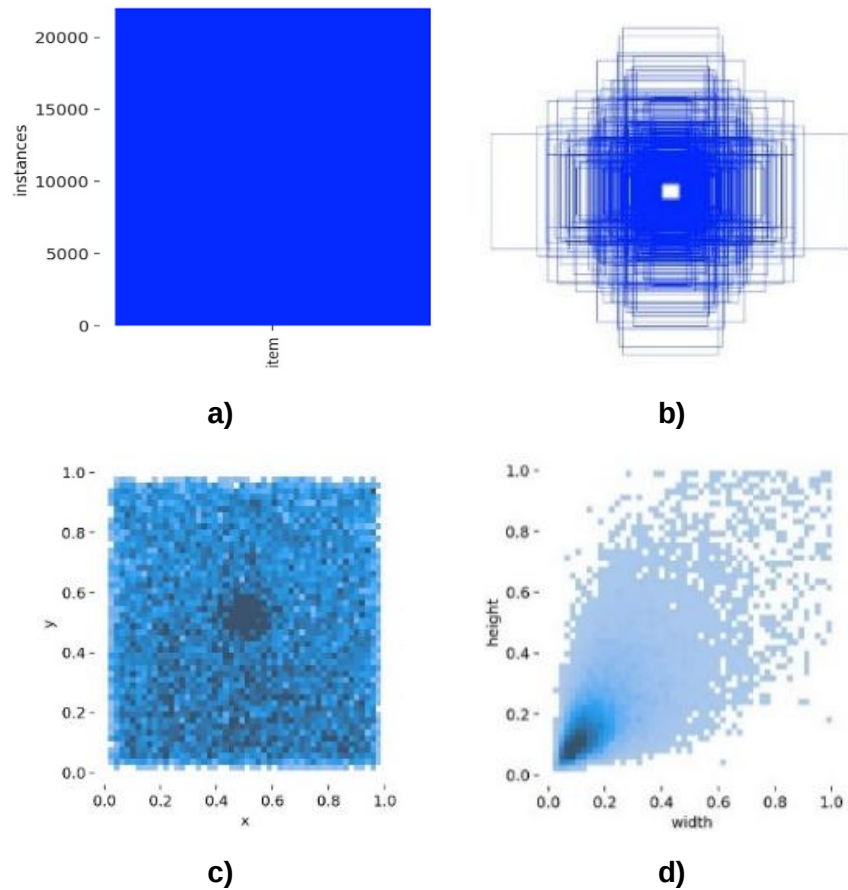


Figura 7.1. Propiedades de las instancias del banco de imágenes propuesto.

El gráfico de barras presentado en el inciso (a) ilustra la distribución del conjunto de datos, el cual consta de un total de 27,397 instancias pertenecientes a una única clase. Al acotar el problema a una categoría, se busca que la red neuronal optimice su capacidad de generalización para modelar la alta variabilidad, derivada de las transformaciones morfológicas y fenotípicas que exhiben las hojas en entornos no controlados.

El inciso (b), ilustra la superposición geométrica de las cajas delimitadoras que se encuentran centradas en un centro común, lo que permite evaluar de forma directa la variabilidad de formas y tamaños que predomina en el conjunto de datos.

El inciso (c), muestra la distribución espacial de las formas de dichas cajas; el área más oscura denota que la mayoría de las estructuras foliares poseen una geometría cercana a la de un cuadrado, dicha área se ubica alrededor de las coordenadas x,y (0.5,0.5).

El inciso (d), expone el gráfico de dispersión correspondiente a las magnitudes de las cajas, donde la zona azul oscura denota la mayor densidad de frecuencia de los objetos. En este último se aprecia una acumulación crítica en los intervalos [0.0, 0.2] tanto para el eje de anchura (*width*) como de altura (*height*), lo que demuestra matemáticamente que el conjunto está compuesto predominantemente por instancias de pequeñas dimensiones.

b) Hiperparámetros

La Tabla 7.2 (pp. 73 – 74), muestra los hiperparámetros definidos para el entrenamiento y la evaluación de las arquitecturas evaluadas en los diferentes escenarios experimentales.

Tabla 7.2. Hiperparámetros genéricos de la familia *Yolo*.

Atributo	Descripción	Valor
<i>Batch</i>	Número de imágenes por lote.	47
<i>Close_mosaic</i>	Deshabilita el aumento de datos mosaico 50 épocas antes de que finalice el entrenamiento.	50
<i>Epochs</i>	Número de iteraciones.	500
<i>Imgsh</i>	Tamaño de la imagen.	416
<i>Optimizer</i>	Función de optimización.	SGD
<i>Patience</i>	Detiene el proceso de aprendizaje cuando no existen cambios significativos en la pérdida.	25
<i>Single_cls</i>	Habilita la clasificación binaria.	True
<i>Cos_lr</i>	Contribuye a una mejor convergencia del modelo.	True
<i>Cache</i>	Mejora los tiempos de entrenamiento.	'Disk'
<i>lr0</i>	Taza de aprendizaje inicial.	0.01
<i>lrf</i>	Taza de aprendizaje final.	0.0001
<i>momentum</i>	Factor de impulso para el optimizador SGD	0.932
<i>weight_decay</i>	Término de regularización que penaliza los pesos grandes para evitar el sobreajuste.	0.0005
<i>warmup_epochs</i>	Número de épocas para el calentamiento de la tasa de aprendizaje, aumentando gradualmente la tasa de aprendizaje desde un valor bajo hasta la tasa de	1.22935

Atributo	Descripción	Valor
	aprendizaje inicial para estabilizar el entrenamiento desde el principio.	
<i>warmup_momentum</i>	Impulso inicial para la fase de calentamiento, ajustándose gradualmente al impulso establecido durante el periodo de calentamiento.	0.1525
<i>box</i>	Peso del componente de pérdida, que influye en el énfasis que se pone en predecir con precisión las coordenadas de la caja delimitadora.	18.27875
<i>cls</i>	Peso de la pérdida de clasificación, que afecta a la importancia de la predicción correcta de la clase en relación con otros componentes.	1.32899
<i>dfl</i>	Peso de la pérdida focal de distribución, utilizado en algunas versiones de YOLO para una clasificación refinada.	0.56016
<i>hsv_h</i>	Modelo de color <i>hsv</i> (aumento de matiz)	0.0
<i>hsv_s</i>	Modelo de color <i>hsv</i> (aumento de saturación)	0.0
<i>hsv_v</i>	Modelo de color <i>hsv</i> (aumento de valor)	0.0
<i>degrees</i>	Rotación de la imagen (valor entero expresado en grados).	0.0
<i>translate</i>	Desplazamiento del objetivo.	0.12431
<i>scale</i>	Agrandamiento del objetivo.	0.07643
<i>shear</i>	Recorte del objeto.	0.0
<i>perspective</i>	Cambio de perspectiva.	0.0
<i>flipud</i>	Efecto espejo arriba y abajo (valor).	0.63
<i>fliplr</i>	Efecto espejo izquierda y derecha.	0.63
<i>mosaic</i>	Crea una nueva imagen al fusionar cuatro imágenes del <i>dataset</i> en una.	0.63
<i>mixup</i>	Combinación de dos o más técnicas de aumento de datos.	0.0

Todos los experimentos se llevaron a cabo en la plataforma de (GoogleColab, s.f), utilizando un entorno virtual equipado con: *Intel(R) Xeon(R) CPU @ 2.20GHz*, una *GPU NVIDIA T4 (15 GBytes VRam)*, *12.7 GBytes RAM* del sistema y *78 GBytes* de disco duro.

7.2 Estudio de ablaciones con la arquitectura de *Yolov11n*

Se integró un plan de ocho casos de experimentación que incluye a sustitución de la función de pérdida original *CIoU* que estima la ubicación del objeto de interés en la imagen por *WIoUv3*. La versión mejorada incorpora la distancia de atención entre centroides y la relación espacial de las cajas delimitadoras, evaluando dinámicamente la calidad de las

cajas de anclaje (ABs) para inferir una localización precisa del objetivo al introducir un coeficiente focal no monótono (Wang *et al.*, 2023) (Tong *et al.*, 2023).

Asimismo, en el módulo (*backbone*) de la arquitectura base *Yolov11n*, se ejecutan los reemplazos estructurales en los índices siete (*Conv*), ocho (*C3k2*) y diez (*C2PSA*) correspondientes por los submódulos *MV2Block* y *MobileViTBlock*. Estos componentes reducen la carga de parámetros del modelo e incorporan la fusión de relaciones espaciales locales y globales (Sachin & Mohammad, 2022).

La Tabla 7.3, detalla la configuración del estudio de ablación propuesto para identificar la combinación de mayor desempeño. Esta integración ocurre conjuntamente al sustituir tanto función de regresión (*CloU* o *WIoUv3*) como en el *backbone*, implementado cada uno de los bloques mencionados previamente de manera intercalada, lo que permite medir la inferencia del modelo con los componentes originales y los propuestos.

Tabla 7.3. Casos de experimentación modificando *Yolo11n*.

Núm. Experimento	<i>CloU</i>	<i>WIoUv3</i>	Ablación 1	Ablación 2	Ablación 3
1	✓				
2	✓		✓		
3	✓		✓	✓	
4	✓		✓	✓	✓
5		✓			
6		✓	✓		
7		✓	✓	✓	
8		✓	✓	✓	✓

La Tabla 7.4 resume los resultados alcanzados en las ocho configuraciones antes especificadas usando el conjunto de validación. Para facilitar la interpretación visual a lo largo del reporte experimental, los valores numéricos en color azul expresan los mejores resultados, mientras que el color rojo señala un rendimiento inferior.

Tabla 7.4. Resultados alcanzados con las ocho configuraciones del modelo.

Núm. Experimento	Parámetros	<i>GFlops</i>	<i>Precision</i>	<i>Recall</i>	<i>F1</i>	<i>mAP @50</i>	<i>mAP @50-95</i>
1	2,582,347	6.3	84.60%	79.70%	82.28%	88.40%	65.20%
2	2,289,069	6.1	83.80%	79.40%	81.53%	88.10%	65.20%
3	1,959,693	5.8	84.20%	80.10%	82.09%	88.30%	65.30%
4	1,891,501	5.8	85.00%	79.90%	82.37%	88.50%	65.60%
5	2,582,347	6.3	82.60%	79.10%	80.81%	87.60%	63.80%
6	2,289,069	6.1	84.80%	78.80%	81.68%	88.00%	64.80%
7	1,959,693	5.8	85.50%	78.80%	82.00%	88.00%	64.80%
8	1,891,501	5.8	83.80%	81.70%	82.73%	88.50%	65.40%

Al finalizar la evaluación de los ocho escenarios, la variante en el experimento número ocho (*YoloLeaf*) es el candidato aceptable debido a que, las métricas de evaluación *Recall* y *F1* son mejores comparados con el resto de casos y el modelo base *Yolov11n*. Adicionalmente, el modelo registró una mayor eficiencia computacional al reducir sensiblemente tanto el número de parámetros del sistema como el volumen de operaciones de punto flotante.

c) Curvas de aprendizaje

Con base en los resultados de la Tabla 7.4 se seleccionó la configuración número ocho (*YoloLeaf*) como la arquitectura final para su evaluación. El entrenamiento de este modelo registró una duración total de 1 h 49 min y 48 s, e incorporó la técnica de parada temprana (*Early Stopping*) con una paciencia (*patience*) de 25 épocas para prevenir el sobreajuste (*overfitting*).

En la Figura 7.2, se muestran cuatro gráficas en donde se aprecia el comportamiento de las funciones el descenso de la pérdida a razón del número de épocas. Los dos gráficos del lado izquierdo (superior e inferior) ilustran la pérdida asociada a la estimación de las coordenadas del rectángulo delimitador que rodea el objetivo. Los gráficos del lado derecho (superior e inferior) expresan la pérdida relacionada al cálculo de probabilidad de pertenencia a la clase del objeto de interés. En breve, se hace la interpretación de cada esquema.

En la gráfica superior izquierda de la Figura 7.2, se observa que la curva de aprendizaje tiende a descender de manera rápida y constante. Esta tendencia refleja una rápida convergencia del optimizador, demostrando que, a medida que avanzan las épocas, el modelo, refina de forma continua la regresión de las coordenadas de las cajas delimitadoras (BBs) y minimiza la penalización por error de ajuste, sin recibir mayores penalizaciones en el proceso. En la gráfica ubicada en el cuadrante inferior izquierdo exhibe de igual manera una trayectoria descendente análoga de pendiente pronunciada, lo que confirma que el modelo logra generalizar el cálculo de los límites espaciales sobre el conjunto de imágenes de validación.

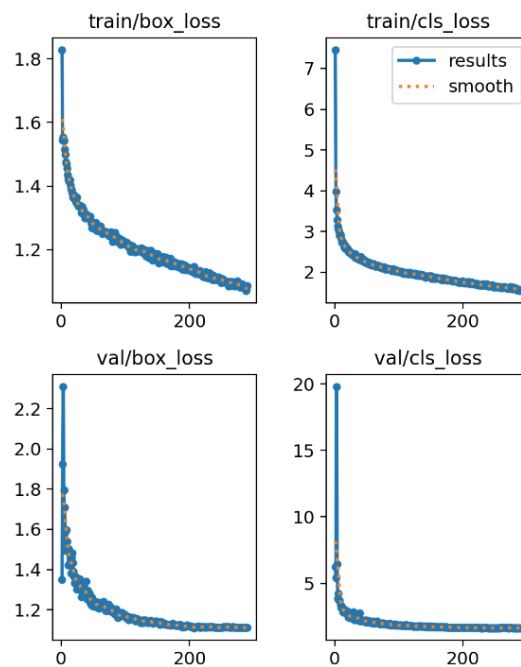


Figura 7.2. Curvas de aprendizaje de *YoloLeaf*.

En el cuadrante superior derecho de la Figura 7.2 se ilustra el progreso en el aprendizaje de la clasificación categórica. La curva exhibe una convergencia rápida y sostenida hacia

cero indicando que el modelo no recibe ningún tipo de castigo y, por ende, aprende a clasificar el conjunto de entrenamiento. En el cuadrante inferior derecho correspondiente a la validación, se distingue un comportamiento constante que va en decremento, la función de pérdida muestra ligeras fluctuaciones al inicio antes de estabilizarse en una trayectoria decreciente uniforme, ratificando la capacidad de discriminación del modelo ante muestras no observadas.

Cabe señalar que el conjunto de entrenamiento cuenta con 22,013 instancias, mientras que la partición de validación dispone de 5,384. Esta disparidad en la escala muestral explica por qué la tasa de convergencia varía entre ambas curvas, dado que una mayor cantidad de objetivos a calcular introduce una mayor variabilidad en la estimación de los gradientes antes de alcanzar el mínimo de pérdida.

Las gráficas de la Figura 7.3 muestran la evolución de las métricas de rendimiento alcanzadas por *YoloLeaf* a lo largo del proceso de entrenamiento.

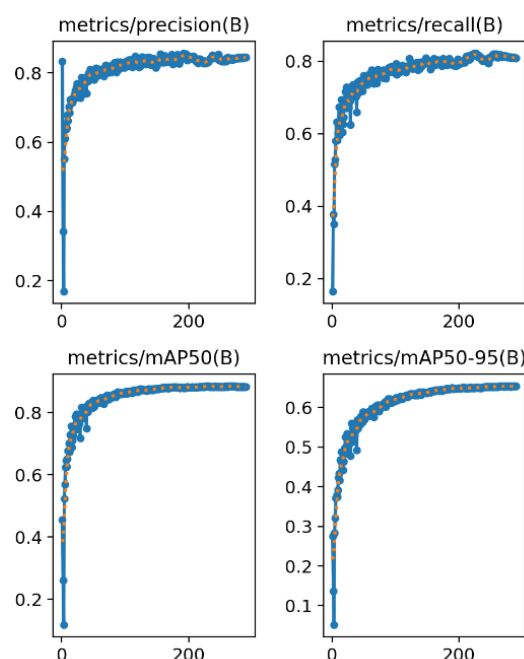


Figura 7.3. Curvas de evaluación de *YoloLeaf*.

En el cuadrante superior izquierdo de la Figura 7.3, se distingue que la curva de precisión muestra una clara tendencia ascendente con fluctuaciones moderadas durante el proceso, lo cual indica que *YoloLeaf* presenta algunas complicaciones para determinar la pertenencia del objeto a dicha clase, pero reduce progresivamente las falsas detecciones (*falsos positivos*). En el cuadrante superior derecho, la curva de sensibilidad (*Recall*) exhibe un comportamiento ascendente similar, reflejando la capacidad del modelo para minimizar las omisiones de detección (*falsos negativos*) e identificar la totalidad de los objetivos presentes.

La gráfica del lado izquierdo inferior (Figura 7.3) presenta un rápido crecimiento de la curva, con perturbaciones mínimas al comienzo, posteriormente se alcanza una estabilización a

medida que las épocas crecen. Esto significa que *YoloLeaf* genera *BBs* cuyas proporciones son cada vez más precisas, cercano a ocupar un área en común del 0.5 (50%) al producirse una intersección con los *BBs* definidos en el *GT*. En el cuadrante inferior derecho, se distingue un comportamiento semejante donde se percibe una aceleración en el crecimiento de la curva acompañado por algunas variaciones que van disminuyendo a lo largo del proceso iterativo. La métrica *mAP50-95*, semejante a la anterior evalúa que las coordenadas de los *BBs* predichas sean cada vez más exactos. En este caso se evalúa que el área en común cuando se realiza una intersección con los *BBs* establecidos en el *GT* incremente gradualmente a partir de 0.50 hasta 0.95 en intervalos de 0.05 (5%).

La Figura 7.4 ilustra la curva *F1-Confidence* de respuesta que involucra la métrica *F1-Score* y un umbral de confianza, ambas variables se encuentran acotadas en el intervalo $[0,1]$. *F1-Score* cuantifica el promedio ponderado de *Precision* y *Recall*; el análisis de esta curva permite determinar el umbral de confianza óptimo que maximiza el equilibrio entre la reducción de falsos positivos y la minimización de falsos negativos para la inferencia del modelo y asegurar un desempeño del modelo aceptable. Permite visualizar el incremento del valor *F1-Score* a media que el grado de confianza crece. En este caso particular se sugiere establecer el grado de confianza igual a 0.449 o 44.90% para obtener un puntaje *F1-Score* igual a 0.83 o 83%.

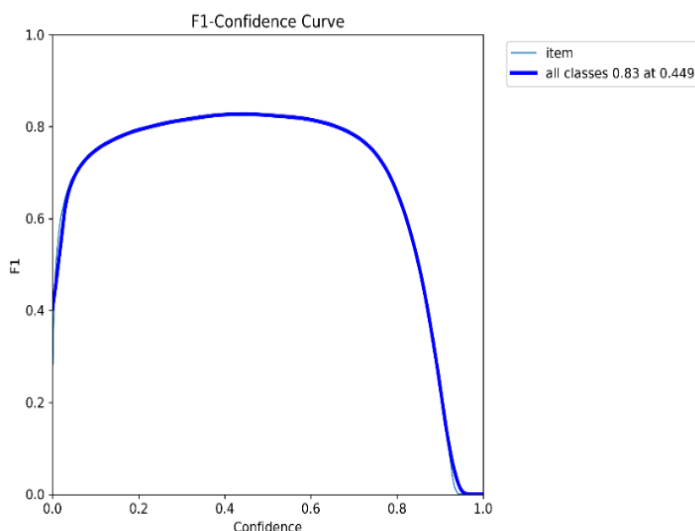


Figura 7.4. Curva *F1-Confidence*.

7.3 Validación cruzada

Dado que la evaluación inicial presentada se basó en una división convencional 80/20, lo que limita la evaluación integral sobre la totalidad del conjunto de datos de referencia (*GT*), se adoptó un esquema de validación cruzada de cinco *folds* ($k = 5$). Este enfoque asegura la exploración completa del dataset al establecer particiones complementarias e independientes.

La Tabla 7.5 muestra la configuración empleada de cinco divisiones donde, cada segmento se encuentra conformado por 80% entrenamiento (E) y 20% para validación (V). Nótese

cómo las proporciones 80/20 se mantienen cuando las distribuciones varían en cada división.

Tabla 7.5. Configuración de la *validación cruzada realizada*.

<i>f</i>	<i>Dataset</i>				
1	V	E	E	E	E
2	E	V	E	E	E
3	E	E	V	E	E
4	E	E	E	V	E
5	E	E	E	E	V

Los resultados presentados en la Tabla 7.6 se obtuvieron mediante el esquema de validación cruzada antes descrito, empleando la misma configuración de hiperparámetros detallada en la Tabla 7.2. El propósito de este experimento es evaluar la estabilidad, consistencia y capacidad de generalización del modelo frente a la variabilidad en la distribución de los subconjuntos de datos a lo largo de los diferentes *folds*.

Tabla 7.6. Equiparación de resultados.

<i>Fold</i>	<i>Precision</i>	<i>Recall</i>	<i>F1</i>	<i>mAP@50</i>	<i>mAP@50-95</i>
1	81.70%	79.30%	80.48%	87.20%	63.70%
2	84.70%	78.80%	81.64%	87.30%	64.60%
3	84.30%	78.50%	81.29%	87.30%	64.30%
4	84.90%	79.40%	82.05%	88.60%	65.40%
5	83.90%	79.60%	81.69%	87.40%	64.30%

En la Tabla 7.7 se presenta el estudio estadístico sobre la variabilidad de los indicadores de evaluación obtenidos. El propósito de esta evaluación es verificar la estabilidad de *YoloLeaf*, asegurando que las puntuaciones no exhiban fluctuaciones atípicas entre las cinco divisiones. Asimismo, este análisis permite confirmar la homogeneidad en la distribución del conjunto de imágenes y descarta la presencia de sesgos o anomalías en la constitución de las particiones.

Tabla 7.7. Análisis estadístico.

	<i>Precision</i>	<i>Recall</i>	<i>F1</i>	<i>mAP@50</i>	<i>mAP@50-95</i>
μ	83.90	79.12	81.43	87.56	64.46
σ^2	1.32	0.16	0.27	0.26	0.30
σ	1.15	0.40	0.51	0.50	0.54

Como se observa en la Tabla 7.7, se presenta el estudio estadístico sobre la variabilidad de los indicadores de evaluación obtenidos con *Recall*, *F1*, *mAP@50* y *mAP@50-95*; la varianza y la desviación estándar son mínimas ya que, no superan la unidad de cambio mientras que, el indicador *Precision* supera por poco la unidad en cada caso.

La Tabla 7.8 presenta la comparación de métricas y características del modelo original *Yolov11n* (modelo 1) y el modelo *YoloLeaf* (modelo 8). Adicionalmente, se incluye el tamaño del modelo expresado en *Mbytes*.

Tabla 7.8. Comparación de características y métricas I.

Modelo	Tamaño (Mbytes)	Parámetros	GFlops	Precision	Recall	F1	mAP @50	mAP @50-95
1	5.20	2,582,347	6.3	84.60%	79.70%	82.28%	88.40%	65.20%
8	3.90	1,891,501	5.8	83.80%	81.70%	82.73%	88.50%	65.40%

En conclusión, *YoloLeaf* (modelo ocho) presenta un desempeño competitivo respecto al modelo base *Yolov11n* (modelo uno), al registrar una mejora en cuatro métricas. Además de registrar una reducción del tamaño de almacenamiento, el número de parámetros de entrenamiento y las operaciones de punto flotante.

7.4 Comparación de *YoloLeaf* contra *Yolov12n* y *Yolov11n*

A inicios del año 2025 (*Ultralytics, s.f*) publicó una nueva versión denominada *Yolov12* que se basa en la arquitectura de su predecesora (*Yolov11*). La principal innovación de *Yolov12* es la incorporación de ViTs (*Tian et al., 2025*).

En la Tabla 7.9 se muestran los resultados obtenidos con *Yolov12n*, implementando el entrenamiento tradicional 80/20, con los mismos hiperparámetros descritos anteriormente (Tablas 7.1 y 7.2) y sin cambiar la función de pérdida original (*CIoU*). Asimismo, se añaden los resultados de *Yolov11n* y *YoloLeaf* para realizar una comparación cuantitativa.

Tabla 7.9. Comparación de características y métricas II.

Modelo	Tamaño (Mbytes)	Parámetros	GFlops	Precision	Recall	F1	mAP @50	mAP @50-95
<i>Yolov12n</i>	5.25	2,556,923	6.3	85.20%	78.30%	81.60%	87.60%	64.70%
<i>Yolov11n</i>	5.20	2,582,347	6.3	84.60%	79.70%	82.28%	88.40%	65.20%
<i>YoloLeaf</i>	3.90	1,891,501	5.8	83.80%	81.70%	82.73%	88.50%	65.40%

En las Tablas 7.10 y 7.11, se analiza el costo beneficio entre *YoloLeaf* contra *Yolov11n* y *Yolov12n* empleando la ecuación 7.1:

$$\Delta \% = \frac{V_n - V_a}{V_a} (100) \quad \text{Ecuación 7.1}$$

donde V_n es el valor nuevo (*YoloLeaf*); V_a es el valor antiguo (*Yolov12n* o *Yolov11n*).

Este cálculo de cambio porcentual se implementa en el resto de este capítulo para hacer una comparación mediante las cuatro métricas protagonistas (*Precision*, *Recall*, *mAP@50*, *mAP@50-95*) y, determinar si existe un avance positivo o negativo entre los modelos bajo análisis.

El sentido de las flechas en las Tablas 7.10 y 7.11 representa la tendencia ascendente o descendente estimada, según corresponda a cada caso.

Tabla 7.10. Comparación del costo beneficio con tendencia descendente.

Característica		YoloLeaf Vs Yolov11n	YoloLeaf Vs Yolov12n
Tamaño	↓	25.00%	25.71%
Núm. de parámetros	↓	26.75%	26.02%
GFlops	↓	7.93%	7.93%

En este caso los datos obtenidos son favorables debido a que, el tamaño de almacenamiento, el número de parámetros y la complejidad del cálculo de operaciones flotante reflejan reducciones importantes de dos dígitos y un dígito respectivamente, lo que significa que el modelo *YoloLeaf* requiere de menos recursos computacionales para su funcionamiento.

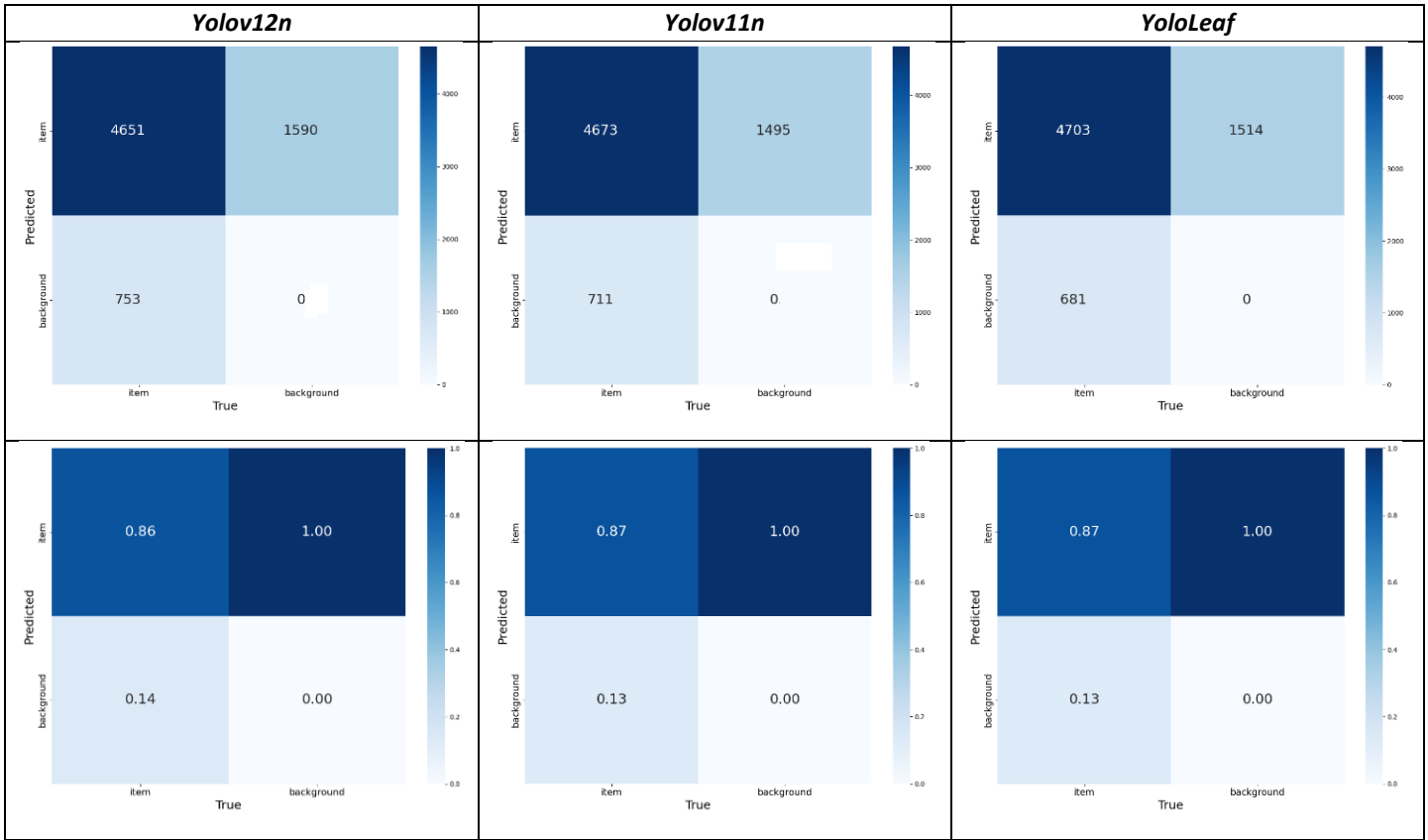
Tabla 7.11. Comparación del costo beneficio con tendencia ascendente.

Métricas		YoloLeaf Vs Yolov11n	YoloLeaf Vs Yolov12n
<i>Precision</i>	↑	-0.94%	-1.64%
<i>Recall</i>	↑	2.50%	4.34%
<i>F1</i>	↑	0.54%	1.38%
<i>mAP@50</i>	↑	0.11%	1.02%
<i>mAP@50-95</i>	↑	0.30%	1.08%

En esta situación al cuantificar el cambio porcentual con respecto a *Yolov11n* y *Yolov12n*, se destaca el incremento en la sensibilidad (*Recall*) de 2.50% y 4.34%, respectivamente. Esto indica que *YoloLeaf* es menos propenso a omitir objetos de estudio pertenecientes a la clase de interés (falsos negativos). Sin embargo, la contracción del 0.94% y 1.64% de *Precision* respectivamente, indica que el modelo es más susceptible a clasificar erróneamente otros objetivos que no pertenecen a la clase (falsos positivos).

En la Tabla 7.12 se analizan seis matrices de confusión correspondientes a los modelos bajo estudio. En la primera fila los valores numéricos son expresados en términos absolutos, mientras que en la segunda fila las cantidades son expresadas en términos porcentuales.

Tabla 7.12. Comparación de matrices de confusión.



De acuerdo con los resultados de cada matriz de confusión, *Yolov12n* presenta un desempeño inferior al de *Yolov11n* y *YoloLeaf*, al registrar cifras elevadas de falsos positivos (1,590) y falsos negativos (753), siendo evidente una mayor probabilidad de inferencia incorrecta.

Por su parte, *Yolov11n* registra 1,495, que es la menor cantidad de falsos positivos en los tres casos, lo que demuestra que es menos susceptible a tener errores de este tipo. Sin embargo, los 711 de falsos negativos señala que existe una posibilidad de no detectar correctamente los objetivos de interés presentes en la imagen.

YoloLeaf registra la menor cifra de falsos negativos (681) de los tres modelos, lo que demuestra una menor probabilidad de omitir objetos de interés presentes en la imagen. No obstante, los 1,514 falsos positivos reflejan una mayor tendencia a inferir erróneamente que otros objetos ajenos pertenezcan a la clase de interés respecto a *Yolov11n*, pero mejorando lo reportado por *Yolov12n*.

7.5 Análisis cuantitativo por casos, YoloLeaf contra YOLOv12n y YOLOv11n

Esta prueba tiene como objetivo evaluar el desempeño de los modelos *YoloLeaf*, *YOLOv12n* y *YOLOv11n*, en la localización de hojas de jitomate en imágenes que presentan escenarios de diversos entornos de campo abierto. Los entornos evaluados se encuentran categorizados en 10 casos de estudio. La Figura 7.5 presenta la distribución de las 5,384 instancias etiquetadas que componen el conjunto de validación. Como se observa, el 60.10% de las instancias provienen de capturas realizadas bajo condiciones ambientales adversas, mientras que el 39.90% restante corresponde a escenarios de menor complejidad.

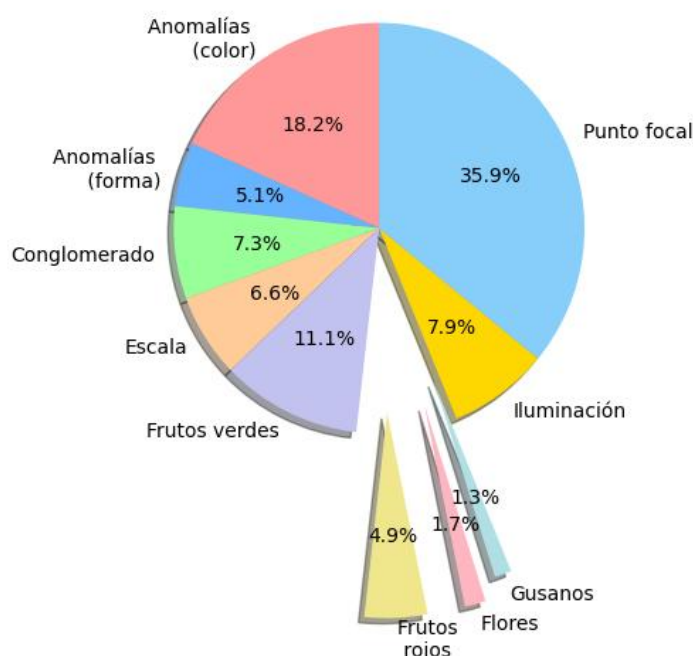


Figura 7.5. Distribución de instancias.

La Tabla 7.13 (pp. 83 - 84) detalla la organización y comparativa de las predicciones de los tres modelos evaluados, así como la distribución del número de imágenes e instancias (cajas delimitadoras) consideradas.

Tabla 7.13. Análisis del desempeño de los tres modelos.

Anomalías (color)						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	81.30%	78.20%	82.90%	62.10%	981	116
Yolov11n	83.70%	76.60%	83.50%	62.90%		
YoloLeaf	86.20%	76.00%	84.20%	63.40%		
Anomalías (forma)						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	82.30%	61.60%	78.10%	48.20%	272	24
Yolov11n	73.00%	70.20%	76.70%	48.90%		
YoloLeaf	78.20%	67.10%	75.10%	47.20%		

Conglomerado						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	81.30%	68.70%	80.10%	54.50%	392	25
Yolov11n	73.00%	77.90%	82.10%	56.80%		
YoloLeaf	76.50%	77.20%	83.70%	57.70%		
Escala						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	85.60%	84.80%	89.10%	62.30%	355	20
Yolov11n	90.70%	79.90%	91.00%	63.70%		
YoloLeaf	88.60%	85.60%	91.30%	61.90%		
Frutos verdes						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	75.20%	73.70%	79.70%	56.00%	596	48
Yolov11n	76.30%	79.30%	83.00%	57.40%		
YoloLeaf	78.30%	76.50%	82.40%	57.20%		
Frutos Rojos						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	82.60%	73.30%	83.30%	58.00%	266	22
Yolov11n	85.50%	73.70%	84.50%	58.00%		
YoloLeaf	76.20%	82.00%	85.50	58.60%		
Flores						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	71.90%	83.10%	84.70%	59.90%	89	15
Yolov11n	82.70%	82.00%	87.90%	62.70%		
YoloLeaf	89.00%	83.70%	87.90%	61.40%		
Gusanos						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	82.70%	76.40%	85.40%	61.50%	72	11
Yolov11n	90.00%	66.70%	82.30%	59.10%		
YoloLeaf	84.40%	75.00%	82.60%	60.80%		
Iluminación						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	89.50%	80.00%	90.90%	67.80%	426	44
Yolov11n	86.70%	79.50%	89.90%	66.50%		
YoloLeaf	83.40%	82.60%	90.30%	67.50%		
Punto focal						
Modelo	Precision	Recall	mAP@50	mAP@50-95	Instancias	Imágenes
Yolov12n	88.60%	87.30%	94.10%	73.60%	1,935	275
Yolov11n	90.30%	86.60%	94.40%	73.50%		
YoloLeaf	88.00%	88.60%	94.40%	73.90%		

Para una inspección visual de los puntajes recabados en los tres modelos estudiados (ver Tabla 7.13) en la Figura 7.6 se ilustran cuatro gráficos de telarañas en función de las métricas de *Recall* y *Precision*.

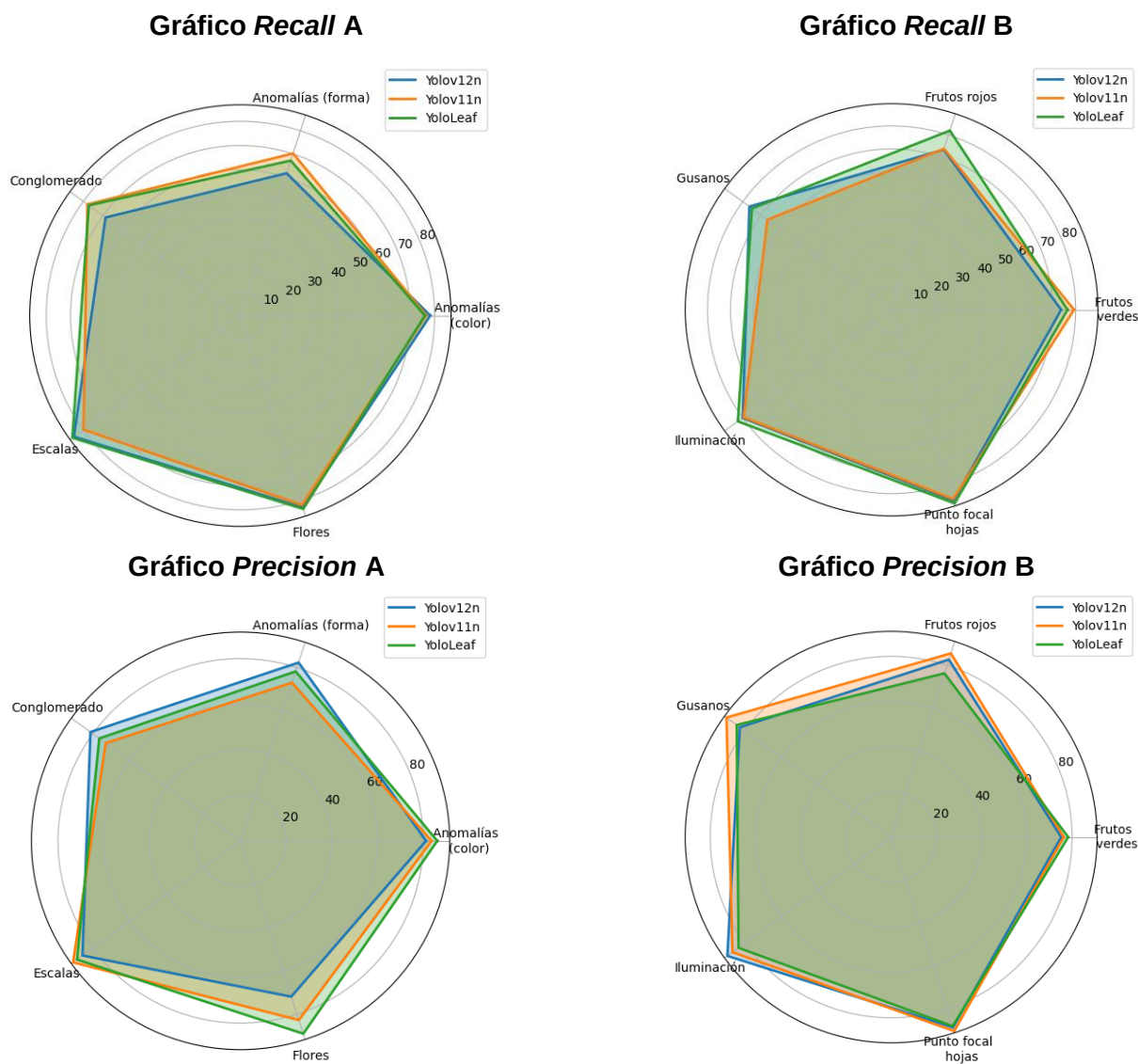


Figura 7.6. Análisis de las métricas *Recall* y *Precision*.

Como se aprecia en las gráficas, *YoloLeaf* logró una mayor recuperación de hojas existentes en imágenes donde figuran frutos rojos, seguido de escenarios con variaciones de iluminación sobre el follaje. En los casos de estudio correspondientes a "escalas", "punto focal" y "flores", el modelo obtuvo un rendimiento ligeramente superior frente a *Yolov12n* y *Yolov11n*. Por el contrario, los únicos escenarios donde *YoloLeaf* registró un desempeño ligeramente inferior correspondieron a "conglomerados" y "gusanos", en comparación con *Yolov11n* y *Yolov12n* respectivamente.

En lo relativo a la *Precision*, se observa que *YoloLeaf* presenta menos casos de falsos positivos en imágenes que incluyen flores de la planta de jitomate. Además, en las circunstancias de "anomalías de color" y "frutos verdes" se distingue un mínimo progreso con respecto a *Yolov12n* y *Yolov11n*.

La Tabla 7.14 registra el conteo de los casos donde cada modelo logró un puntaje mayor comparado con el resto, en términos de las métricas *Precision* y *Recall*.

Tabla 7.14. Conteo del desempeño superior.

Modelo	<i>Precision</i>	<i>Recall</i>
<i>Yolov12n</i>	3 / 10	2 / 10
<i>Yolov11n</i>	4 / 10	3 / 10
<i>YoloLeaf</i>	3 / 10	5 / 10

YoloLeaf destaca por su inferencia en la recuperación de hojas en la mitad de los casos analizados con respecto a *Yolov12n* y *Yolov11n*. Sin embargo, la probabilidad de falsos positivos es igual que *Yolov12n*.

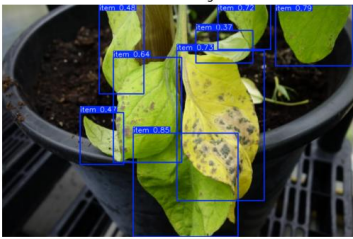
7.6 Pruebas de localización con hojas de jitomate

En las Tablas 7.15 a 7.27 se presentan 13 casos de estudio orientados al análisis cualitativo y cuantitativo de las inferencias generadas por las tres arquitecturas evaluadas en la detección de láminas foliares de jitomate. La evaluación tiene como propósito examinar la capacidad de generalización de cada modelo al ser evaluado con imágenes independientes que no formaron parte del conjunto de datos empleado en las fases previas de entrenamiento y prueba.

- **Oclusiones**

En la imagen (ver Tabla 7.15) se aprecian hojas anómalas del tipo papa de la planta de jitomate que presentan diferentes tipos de solapamiento y de coloración. En la parte central hay una agrupación de tres láminas foliares, dos de ellas exhiben visibilidad reducida debido a su orientación relativa, mientras que la tercera expone solo una fracción de su superficie por la superposición de las dos primeras. Al fondo de la imagen, otras láminas foliares se aprecian parcialmente debido al margen de captura del sensor. Las características principales de la imagen original son: resolución de 933×657 píxeles y, el tipo de formato es *png*.

Tabla 7.15. Evaluación cualitativa y cuantitativa de localización (oclusiones).

<i>Yolov12n</i>		<i>Yolov11n</i>		<i>YoloLeaf</i>	
Number of Bounding boxes:7		Number of Bounding Boxes:5		Number of Bounding Boxes:8	
					
Contador:	7/6	5/6		8/6	
Precision:	74.70%	71.80%		71.10%	
Recall:	98.60%	83.30%		99.99%	
mAP50:	95.55%	88.70%		92.70%	
mAP50-95:	56.20%	43.60%		60.50%	

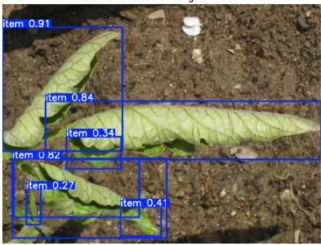
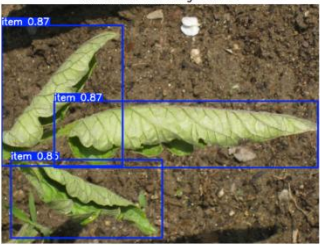
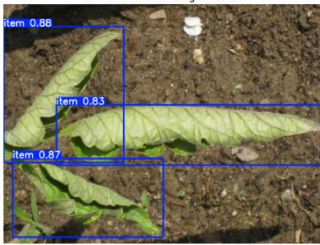
Tanto *Yolov11n* como *Yolov12n* presentaron problemas de inferencia para demarcar correctamente las láminas centrales debido al traslape y la forma que adoptan. No obstante, se observó un mejor desempeño del modelo *YoloLeaf*, ya que contempla correctamente el

área de las hojas (99.99% de *Recall*) y un mejor trazado de los *BBs* (60.50% de *mAP@50-95*).

- **Anomalía I, cambio en la forma**

En este caso de estudio (ver Tabla 7.16), se analiza la detección sobre láminas foliares que exhiben alteraciones morfológicas severas inducidas por factores de estrés no determinados, tales como fitopatologías o déficit hídrico, etc. Las características principales de la imagen original son: dimensiones de 1,024×768 píxeles y, la extensión del archivo corresponde a *webp*.

Tabla 7.16. Evaluación cualitativa y cuantitativa de localización (cambio en la forma).

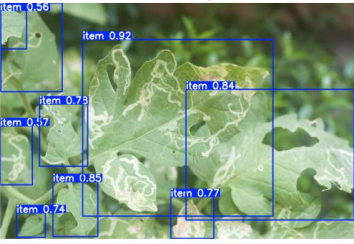
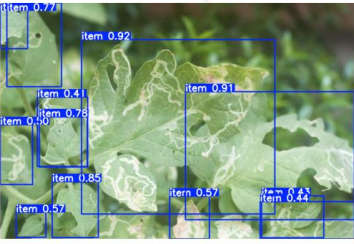
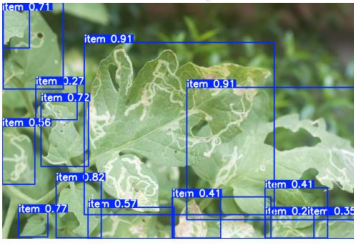
Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:7		Number of Bounding Boxes:3		Number of Bounding Boxes:3	
					
Counter:	7/3	3/3		3/3	
Precision:	97.80%	98.10%		98.00%	
Recall:	99.00%	99.99%		99.99%	
mAP50:	99.50%	99.50%		99.50%	
mAP50-95:	85.20%	80.90%		83.10%	

En cuanto al proceso de inferencia, *Yolov12n* presentó imprecisiones en la delimitación de las regiones afectadas debido a la generación de cajas delimitadoras redundantes. Sin embargo, los puntajes 97.80% de *Precision*, 99.00% de *Recall*, 99.50% de *mAP@50* y 85.20% de *mAP@50-95* confirman que las afecciones son mínimas. En contraste, *YoloLeaf* eliminó la duplicación de cajas delimitadoras, alcanzando valores de 98.00% de *Precision*, 99.99% de *Recall*, 99.50% de *mAP@50* y 83.10% de *mAP@50-95*, incluso el *Recall* es levemente superior a la de a *Yolov12n*.

- **Anomalía II, daños por plaga**

La siguiente alteración en las hojas del tipo sierra (ver Tabla 7.17) es causada por la plaga del minador que deja una serie de lesiones considerables sobre la superficie, afectando principalmente la textura. Adicionalmente, se nota sobre la lámina foliar diversas rupturas lo que causa un cambio notorio en el patrón de la forma del objeto. Las características principales de la imagen original son: tamaño de 1,500×966 píxeles y el formato es *jpg*.

Tabla 7.17. Evaluación cualitativa y cuantitativa de localización (daño por plaga).

Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:9		Number of Bounding Boxes:12		Number of Bounding Boxes:15	
					
Counter:	9/12	12/12		15/12	
Precision:	75.10%	84.10%		91.70%	
Recall:	75.00%	88.20%		91.70%	
mAP50:	86.80%	90.90%		94.80%	
mAP50-95:	64.00%	63.50%		70.30%	

YoloLeaf marcó los mejores puntajes comparado contra Yolov12n y Yolov11n. Las métricas de clasificación (*Precision* y *Recall*) son de 91.70% lo que indica una menor probabilidad de cometer falsos positivos y falsos negativos. En cuanto a los indicadores que miden la calidad de los *BBs* corresponden a 91.70% y 70.30% de *mAP@50* y *mAP@50-95* respectivamente, lo que señala una delimitación precisa del área de interés afectada.

- **Agrupación de hojas sanas**

En este escenario (ver Tabla 7.18), se observa una agrupación de hojas del tipo sierra y algunos jitomates en proceso de maduración. Se nota una variedad de orientaciones, oclusiones y tamaños de las láminas foliares, una de ellas muestra el envés debido al dobles. Las características principales de la imagen original son: proporciones de 667×1,000 píxeles y, formato *webp*.

Tabla 7.18. Evaluación cualitativa y cuantitativa de localización (agrupación de hojas sanas).

Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:30		Number of Bounding Boxes:31		Number of Bounding Boxes:32	
					
Counter:	30/31	31/31		32/31	
Precision:	93.40%	96.70%		92.30%	
Recall:	91.00%	94.80%		93.50%	
mAP50:	97.40%	95.90%		96.10%	
mAP50-95:	75.20%	75.80%		72.10%	

Yolov11n registró un buen desempeño ya que su inferencia muestra ser certero al marcar 96.70% de *Precision* y 94.80% de *Recall*, seguidamente de *YoloLeaf* al registrar 93.50% de *Recall* y 92.30% de *Precision* y, *Yolov12n* al reportar 93.40% de *Precision* y 91.00% de *Recall*. El mejor trazado de las instancias es *Yolov11n* ya que la métrica *mAP@50-95* es de 75.80%, la más alta de los tres modelos.

- **Agrupación de hojas anómalas**

El ejemplar de la Tabla 7.19 expone un conglomerado de láminas foliares del tipo papa y sierra de la planta de jitomate que presentan variaciones de tamaños, coloración (anomalía biológica), oclusiones y de forma. El tamaño de la imagen original es de 1,080×1,440 píxeles y, el tipo de formato es *webp*.

Tabla 7.19. Evaluación cualitativa y cuantitativa de localización (agrupación de hojas anómalas).

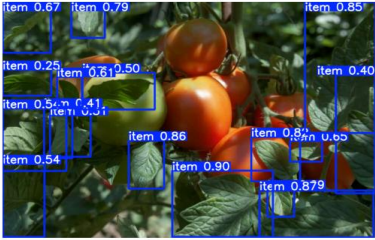
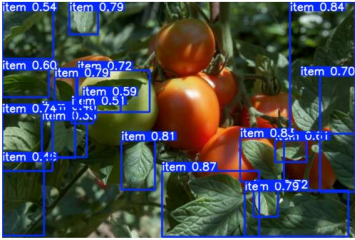
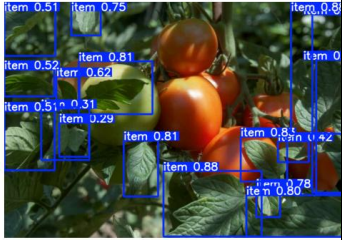
<i>Yolov12n</i>		<i>Yolov11n</i>		<i>YoloLeaf</i>	
Number of Bounding Boxes:36 		Number of Bounding Boxes:29 		Number of Bounding Boxes:35 	
Counter:	36/46	29/46		35/46	
Precision:	87.40%	91.90%		96.79%	
Recall:	67.40%	54.30%		68.90%	
mAP50:	80.10%	76.70%		82.00%	
mAP50-95:	53.00%	49.70%		52.30%	

Se observó que la red propuesta alcanzó los mejores puntajes en las siguientes tres métricas: *Precision* (96.79%), *Recall* (68.90%) y *mAP@50* (82.00%) demostrando un mejor desempeño contra *Yolov12n* al lograr 87.40% de *Precision*, 67.40% de *Recall* y 80.10% de *mAP@50*. Con respecto a *Yolov11n* se reportó 91.90% de *Precision*, 54.30% de *Recall* y 76.70% de *mAP@50*.

- **Iluminación irregular I**

El siguiente escenario (ver Tabla 7.20) considera un entorno complejo ya que, debido a la iluminación irregular se originan sombras que inciden sobre un conjunto de hojas aparentemente sanas, lo que trae como consecuencia pérdida y/o alteración de información relevante como bordes, nervaduras y cambios de coloración en la superficie foliar. Las dimensiones de la imagen original son de 1,000×665 píxeles y, su extensión es *jpg*.

Tabla 7.20. Evaluación cualitativa y cuantitativa de localización (iluminación irregular I).

Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:17		Number of Bounding Boxes:19		Number of Bounding Boxes:17	
					
Counter:	17/17		19/17		17/17
Precision:	92.10%		87.30%		83.20%
Recall:	70.60%		80.90%		87.50%
mAP50:	89.50%		91.70%		89.60%
mAP50-95:	57.50%		60.70%		58.30%

Yolov12n reportó 92.10% de *Precision* pero, 70.60% de *Recall*, lo que significa que algunas áreas están siendo omitidas por el modelo. En cuanto a la calidad del trazado de los BBs se logró 89.50% de *mAP@50* y 57.50% de *mAP@50-95*.

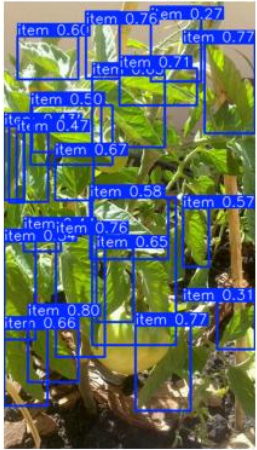
Yolov11n tiene los mejores resultados del cálculo de las instancias delimitadoras al obtener 91.70% *mAP@50* y 60.70% *mAP@50-95*. Además, un avance favorable en el *Recall* de 80.90% pero, la métrica *Precision* es de 87.30%, menor que *Yolov12n*.

YoloLeaf logró un *Recall* de 87.50% superando a *Yolov12n* y *Yolov11n* en este indicador, lo que confirma una mayor precisión de cobertura en regiones donde yacen las hojas.

- **Iluminación irregular II**

El siguiente análisis (ver Tabla 7.21), muestra un escenario complejo a causa de, la existencia de una agrupación de hojas del tipo papa. En algunos ejemplares se percibe de manera parcial la superficie principal y en otros se observa el envés de la lámina foliar. Por otra parte, el tipo de iluminación que incide sobre la planta proyecta sombras en las hojas. La resolución de la imagen original es de 360×640 píxeles y, su tipo de archivo es *jpg*.

Tabla 7.21. Evaluación cualitativa y cuantitativa de localización (Iluminación irregular II).

Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:29		Number of Bounding Boxes:16		Number of Bounding Boxes:23	
					
Counter:	29/45		16/45		23/45
Precision:	88.50%		77.50%		70.30%
Recall:	51.30%		55.60%		64.40%
mAP50:	73.70%		70.70%		69.00%
mAP50-95:	39.80%		43.50%		40.40%

Yolov12n registró 88.50% de *Precision* el más alto comparado contra *Yolo11n* y *YoloLeaf* lo que significa una menor probabilidad de errores del tipo falso positivo. No obstante, el 51.30% de *Recall* señala que el modelo está descartando hojas que realmente existen en la imagen (falso negativo). Respecto a la calidad de la delimitación del área de interés se alcanzó 73.70% de *mAP@50* el mejor con respecto *Yolov11n* y la red propuesta.

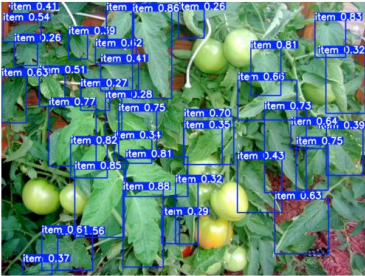
Yolov11n es el modelo que mejor delimita la región de interés, ya que la evaluación en múltiples intervalos marcó 43.50% de *mAP@50-95* pero, en la clasificación de objetivos se registró 77.50% de *Precision* y 55.60% de *Recall*, lo que hace referencia a una mayor probabilidad de falsos positivos y menor probabilidad de falsos negativos con respecto a *Yolov12n*.

YoloLeaf alcanzó 64.40% de *Recall* el mayor porcentaje con respecto a *Yolov12n* y *Yolov11n* lo que apunta a una menor probabilidad de falsos negativos. No obstante, el 70.30% de *Precision* está por debajo de *Yolov11n* y *Yolov12n* lo que significa una mayor probabilidad de falsos positivos.

- **Presencia de jitomates I**

La siguiente imagen en la Tabla 7.22 ilustra un escenario complejo para localizar apropiadamente las hojas de la variante papa y sierra, ya que existe una agrupación importante, lo que ocasiona diversas oclusiones entre ellas. Los jitomates que figuran también son causantes de algunos solapamientos. Además, la distancia es otro factor a considerar, puesto que dependiendo de la escala de las hojas la extracción de información se vuelve desafiante; la imagen original presenta un tamaño de 1,600×1,200 píxeles, en formato *jpg*.

Tabla 7.22. Evaluación cualitativa y cuantitativa de localización (presencia de jitomate I).

Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:36 		Number of Bounding Boxes:37 		Number of Bounding Boxes:49 	
Counter:	36/75		37/75		49/75
Precision:	75.10%		64.40%		75.50%
Recall:	56.00%		48.20%		56.00%
mAP50:	68.70%		64.90%		68.40%
mAP50-95:	41.00%		38.60%		37.80%

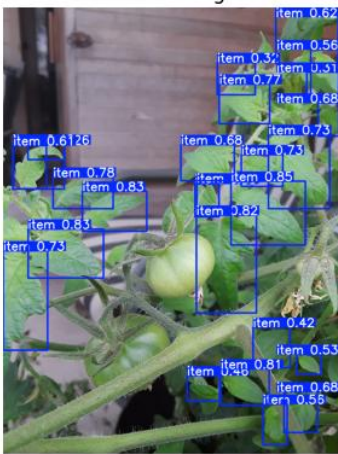
Yolov12n al igual que *YoloLeaf* registraron 56.00% de *Recall* (más alto). No obstante, el modelo planteado arrojó 75.50% de *Precision* lo que significa una menor probabilidad de presentar casos del tipo falsos positivos. *Yolov11n* quedó por detrás al cuantificar un *Precision* de 64.40% y un *Recall* de 48.20%. En estas condiciones adversas, *YoloLeaf* es competente contra *Yolo12n*.

En cuanto a precisión de la demarcación de instancias, *Yolov12n* presentó 41.00% de *mAP@50-95*; el mejor de los tres modelos analizados.

- **Presencia de jitomates II**

En este caso (ver Tabla 7.23), se presenta una captura compleja debido a las múltiples oclusiones entre hojas, tallos y jitomates verdes. Asimismo, la presencia de estos elementos que comparten la característica del color representa un desafío para la correcta inferencia de las hojas; la toma original tiene una dimensión de 1,400×1,867 píxeles, con extensión *jpg*.

Tabla 7.23. Evaluación cualitativa y cuantitativa de localización (presencia de jitomate II).

Yolov12n		Yolov11n	YoloLeaf
Number of Bounding Boxes:26		Number of Bounding Boxes:40	Number of Bounding Boxes:35
			
Counter:	26/35	40/35	35/35
Precision:	89.70%	99.20%	79.20%
Recall:	57.10%	62.90%	76.10%
mAP50:	75.60%	81.30%	85.00%
mAP50-95:	51.30%	51.20%	54.40%

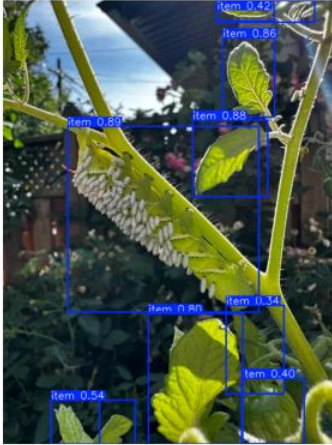
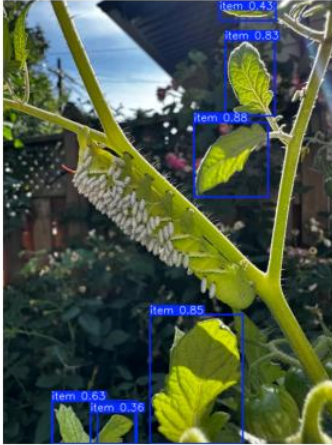
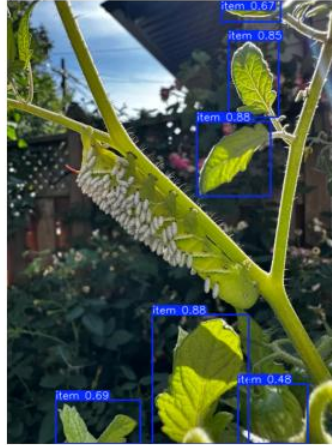
La red propuesta arrojó los mejores resultados al registrar 76.10% de *Recall*, 85.00% de *mAP@50* y 54.40% de *mAP@50-95* lo que significa, una mayor detección de hojas que figuran en la imagen. Sin embargo, la métrica *Precision* de 79.20% señala una mayor probabilidad de registrar casos del tipo falso positivo.

Yolov11n logró mayor *Precision* de 99.20% pero, acompañado de un *Recall* menor de 62.90%, 81.30% de *mAP@50* y 51.20 de *mAP@50-95*. *Yolov12n* obtuvo 89.70% de *Precision*, 57.10% de *Recall*, 75.60% de *mAP@50* y 51.30% de *mAP@50-95*. Aun así, *YoloLeaf* infirió una mayor cobertura donde figuran las láminas foliares ya que, se registraron 85.00% de *mAP@50* y 54.40% de *mAP@50-95*, superiores a *Yolov12n* y *Yolov11n*.

- **Presencia de insecto I**

En esta circunstancia aparte de la presencia de hojas de la planta del jitomate (ver Tabla 7.24), figuran otros elementos como el gusano, tallos y frutos que comparten la característica del color que podrían afectar el resultado de inferencia de los modelos. También, la iluminación ambiental causa cambios de tonalidad y proyección de sombras en la superficie foliar; las proporciones del ejemplar corresponden a 640×853 píxeles, en formato *webp*.

Tabla 7.24. Evaluación cualitativa y cuantitativa de localización (presencia de insecto I).

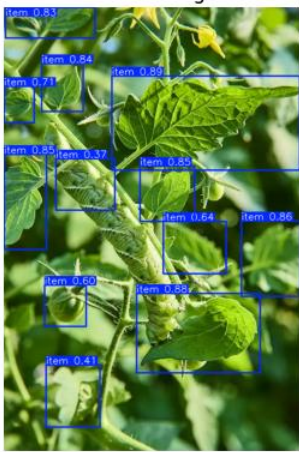
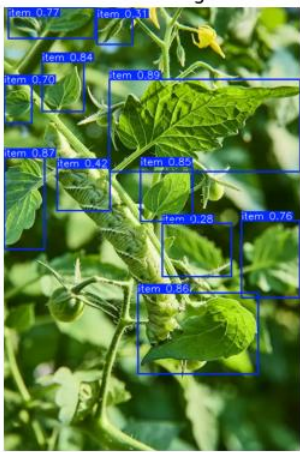
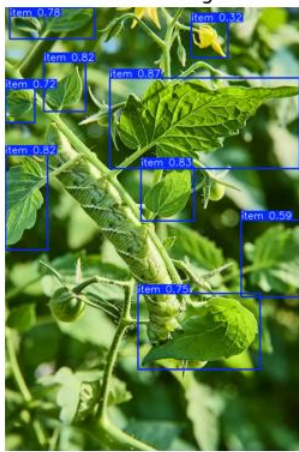
Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:10		Number of Bounding Boxes:7		Number of Bounding Boxes:6	
					
Counter:	10/7	7/7		6/7	
Precision:	71.60%	82.30%		98.30%	
Recall:	85.70%	71.40%		71.40%	
mAP50:	83.70%	80.80%		74.10%	
mAP50-95:	59.60%	66.90%		62.80%	

El modelo *Yolov12n* presentó problemas en su inferencia ya que, el gusano de color verde es detectado como una hoja, causando un porcentaje bajo de 71.60% de *Precision* un dato de imprecisión a considerar. No obstante, es el modelo que reportó un *Recall* de 85.70% mayor que *Yolov11n* (71.40%) y *YoloLeaf* (71.40%) lo que indica que detectó correctamente más cantidad de hojas en la imagen.

- **Presencia de insecto II**

La imagen de la Tabla 7.25 muestra una situación compleja debido a la presencia de varios elementos como: el gusano, tallos y un fruto de jitomate en desarrollo que se asemejan al color de las láminas presentes. Asimismo, el fondo comparte dicha similitud ya que, se conforma por un conglomerado de hojas. Algunas de características de la imagen original son: resolución de 600×900 píxeles y, el tipo de formato es *webp*.

Tabla 7.25. Evaluación cualitativa y cuantitativa de localización (presencia de insecto II).

Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:12		Number of Bounding Boxes:11		Number of Bounding Boxes:9	
					
Counter:	12/9		11/9		9/9
Precision:	98.20%		98.70%		98.90%
Recall:	99.90%		88.90%		88.90%
mAP50:	99.50%		93.30%		95.10%
mAP50-95:	82.80%		78.30		77.10%

Yolov12n reportó los mejores puntajes en tres de las cuatro métricas. El indicador *Precision* es penalizado debido a los tres falsos positivos (fruto verde, parte del gusano y una hoja difuminada del fondo). Sin embargo, el *Recall* es alto (99.90%) ya que se logró localizar correctamente las láminas foliares. De igual modo las dimensiones de los cuadros delimitadores son preciso al registrarse 99.50% y 82.80% de *mAP@50* y *mAP@50-95* respectivamente.

Yolov11n marcó 98.70% de *Precision* por dos casos de falsos positivos (parte del gusano y una flor pequeña). En cuanto a la delimitación de los *Bounding Boxes* se registró 93.90% y 78.30% de *mAP@50* y *mAP@50-95* respectivamente.

YoloLeaf logró 98.90% de *Precision* ya que, sólo se registró un falso positivo (flor de color amarilla), evitando los errores de *Yolov12n* y *Yolov11n* (extremidad del gusano). El *Recall* es menor ya que, una hoja queda excluida (falso negativo).

- **Objetos verdes (macetas)**

La Tabla 7.26 muestra una captura donde figuran un par de macetas de color verde y plantas de jitomate en desarrollo. Además, existe una agrupación de hojas del tipo sierra aparentemente en condiciones saludables. El número de traslapes y la escala afecta a la localización de las láminas foliares. La fuente original presenta las siguientes características: tamaño de 640×480 píxeles y, el tipo de archivo es *.jpg*.

Tabla 7.26. Evaluación cualitativa y cuantitativa de localización (objetos verdes).

Yolov12n		Yolov11n	YoloLeaf
Number of Bounding Boxes:21		Number of Bounding Boxes:20	Number of Bounding Boxes:22
			
Counter:	21/19	20/19	22/19
Precision:	77.80%	85.70%	90.30%
Recall:	78.90%	68.40%	68.40%
mAP50:	82.20%	77.00%	78.90%
mAP50-95:	50.30%	46.50%	47.30%

Los tres modelos, no presentaron errores ante las macetas verdes. Sin embargo, las incidencias ocurren en la agrupación de las hojas principalmente. El mejor desempeño lo registró *Yolov12n* con 78.90% de *Recall*, 82.20% de *mAP@50* y 50.30% de *mAP@50-95*, logrando una mayor precisión de la demarcación de los *BBs* donde existen hojas. *YoloLeaf* arrojó 90.30% de *Precision* pero, en esta situación el *Recall* es de 68.40% igual que *Yolov11n*, una desventaja que indica una mayor probabilidad de casos del tipo falsos negativos en este escenario.

- **Dibujos o impresiones**

En esta imagen (ver Tabla 7.27), aparece una planta de jitomate en fase de crecimiento que tiene varias láminas foliares del tipo sierra. Las hojas presentan diferentes orientaciones, tamaños y algunas oclusiones parciales. Al fondo figuran impresiones de bocetos de hojas. La captura bajo análisis presenta dimensiones de 2,832×4,256 píxeles en formato *jpg*.

Tabla 7.27. Evaluación cualitativa y cuantitativa de localización (dibujos o impresiones).

Yolov12n		Yolov11n		YoloLeaf	
Number of Bounding Boxes:21		Number of Bounding Boxes:20		Number of Bounding Boxes:24	
					
Counter:	21/24		20/24		24/24
Precision:	86.70%		82.80%		83.40%
Recall:	81.80%		83.30%		79.20%
mAP50:	90.40%		90.30%		89.80%
mAP50-95:	60.60%		62.00%		60.50%

La inferencia de *Yolov11n* no muestra problemas ante los bocetos, por consiguiente, el indicador *Recall* mostró un porcentaje mayor de 83.30%. Adicionalmente, la demarcación de las hojas es ligeramente mayor con respecto a los otros modelos en estudio.

Tanto *Yolov12n* y *YoloLeaf* mostraron problemas (falsos positivos) al detectar los dibujos impresos como hojas naturales pertenecientes a la planta.

Finalmente, en la Tabla 7.28 se cuantifica el promedio de las métricas de evaluación de los 13 casos estudiados, así como, el número total de objetivos analizados y el tiempo promedio de inferencia.

Tabla 7.28. Resultados del conjunto de prueba.

Modelo	Precision	Recall	mAP@50	mAP@50-95	Tiempo	Núm. de instancias
Yolo12n	82.30%	65.00%	76.50%	54.30%	10.80 ms	329
Yolo11n	79.90%	60.50%	74.50%	54.40%	8.20 ms	
YoloLeaf	82.90%	67.80%	79.10%	56.30%	9.60 ms	

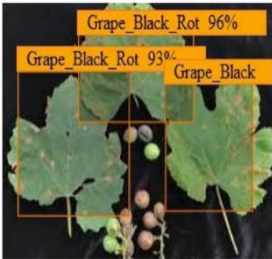
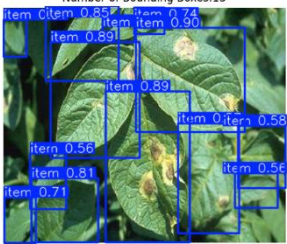
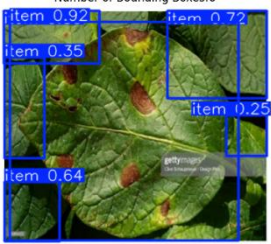
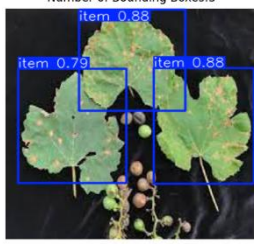
Pese a que los resultados aquí reportados son inferiores a la experimentación del estudio de ablaciones, el modelo propuesto reportó mejores resultados en comparación del modelo base y *Yolov12n*. Esto brinda una perspectiva de la inferencia de *YoloLeaf* en otros entornos, acorde a las características de las imágenes analizadas en esta prueba con otras imágenes fuera de la etapa de experimentación.

7.7 Comparación cualitativa con el estado del arte

En la Tabla 7.29 se realiza una inspección del desempeño de los modelos *YOLO-LeafNet* (Kaur et al., 2025) y *YoloLeaf*. Sólo se hace la evaluación cualitativa ya que, no se cuenta

con el GT y modelo de (Kaur *et al.*, 2025). Las imágenes son recortadas directamente de la publicación; nótese que el punto focal está dirigido hacia las hojas.

Tabla 7.29. Comparación de inferencias *YOLO-LeafNet* contra *YoloLeaf*.

Imágenes originales Kaur <i>et al.</i>, 2025			
<i>YOLO-LeafNet</i> Kaur <i>et al.</i>, 2025			
<i>YoloLeaf</i>	Number of Bounding Boxes:13 	Number of Bounding Boxes:6 	Number of Bounding Boxes:3 
	a)	b)	c)

Como se aprecia en los casos a) y b), el modelo propuesto logró contabilizar una mayor cantidad de hojas con respecto a *YOLO-LeafNet*. En cuanto a c), tanto *YOLO-LeafNet* como *YoloLeaf* logran localizar el mismo número de láminas foliares. Sin embargo, *YoloLeaf* estimó una delimitación de las instancias más precisa, abarcando una mayor región que ocupa el objeto de interés.

7.8 Evaluación cualitativa con otros tipos de hojas del jitomate

De acuerdo con (Rahmadian *et al.*, 2024), en la Tabla 7.30 se detallan las características cualitativas de ocho genotipos de la planta del jitomate, organizados en cuatro variantes de hojas. En la Figura 7.7, se muestran dichos ejemplares localizados por el modelo *YoloLeaf*.

Tabla 7.30. Características cualitativas del tipo de hoja de la planta del jitomate (Rahmadian *et al.*, 2024).

Genotipo	Tipo de hoja
Silimakuta T22	Pimpinellifolium
Raya T23	Potato leaf type
Panei tengahT24	Pimpinellifolium
Purba T25	Standard
Alahan panjang T26	Peruvianum
Magelang T27	Peruvianum
Ratna T28	Peruvianum
Servo F1	Potato leaf type

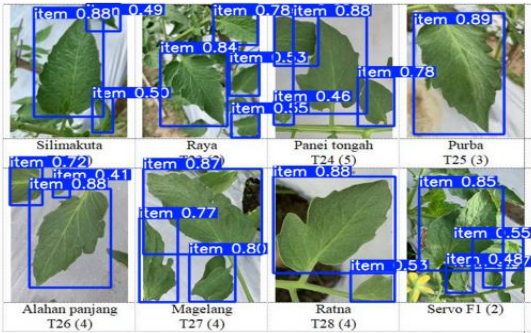


Figura 7.7. Tipos de hojas de la planta del jitomate (Rahmadian *et al.*, 2024).

7.9 Evaluación con otros tipos de cultivos

El objetivo de este análisis es evaluar la generalización del modelo *YoloLeaf* con otros cultivos que presentan hojas con una forma semejante a la hoja de la planta del jitomate, usando las imágenes del conjunto *CWD30*.

Debido a la cantidad del número de imágenes en cada ángulo y semana de desarrollo se determinó analizar: a) el 10% de ejemplares donde el número de capturas es de tres o cuatro dígitos (centenas o millares) y, b) 10 registros donde el número de tomas digitales es de dos dígitos (decenas).

Por ejemplo, en el caso del cultivo de cacahuate, con un ángulo de 0°, semana uno de desarrollo, se tienen 366 imágenes en total (Consultar la Tabla 4.2), por lo que se redondea la cantidad y se evalúan 37 registros aleatoriamente (ver primer columna y primer renglón de la Tabla 7.31); el dígito 319 hace referencia al número de instancias definidas en la captura.

a) Cultivo Cacahuate

La evaluación se efectuó con una muestra conformada de 690 ejemplares que alojan un total de 31,745 cuadros delimitadores. En la Tabla 7.31, se detallan los 19 casos donde se cuantifica la inferencia de *YoloLeaf* acorde al número de semana y captura; los porcentajes mayores se resaltan en color azul.

Tabla 7.31. Evaluación cuantitativa del cacahuate.

Ángulo	Imágenes	Semana	Ángulo 0°		Ángulo 45°		Ángulo 90°	
0°	37 (319)	1	Precision	55.00%	Precision	70.50%	Precision	81.60%
45°	88 (1,102)		Recall	47.50%	Recall	57.40%	Recall	76.00%
90°	32 (597)		mAP50	43.20%	mAP50	63.10%	mAP50	83.80%
157 imágenes (2,018 instancias)			mAP50-90	25.60%	mAP50-90	43.30%	mAP50-90	66.50%
0°	10 (112)	2	Precision	51.20%	Precision	74.60%	Precision	85.30%
45°	91 (1,752)		Recall	43.80%	Recall	65.90%	Recall	71.70%
90°	41 (986)		mAP50	45.20%	mAP50	72.20%	mAP50	82.70%
142 imágenes (2,850 instancias)			mAP50-90	27.60%	mAP50-90	54.50%	mAP50-90	69.90%
0°	10 (228)	3	Precision	66.90%	Precision	76.70%	Precision	81.60%
45°	39 (1,376)		Recall	45.20%	Recall	61.90%	Recall	68.70%
90°	50 (1,679)		mAP50	49.20%	mAP50	72.10%	mAP50	78.50%
99 imágenes 3,283 instancias			mAP50-90	31.30%	mAP50-90	54.40%	mAP50-90	64.20%
-			4	Sin registros	Precision	84.20%	Precision	82.20%
45°	78 (3,520)	Recall			71.40%	Recall	73.00%	
90°	39 (2,208)	mAP50			82.90%	mAP50	81.30%	
117 imágenes (5,728 instancias)		mAP50-90			68.90%	mAP50-90	66.50%	
-			5	Sin registros	Precision	80.80%	Precision	78.20%
45°	56 (4,188)	Recall			67.00%	Recall	65.60%	
90°	31 (2,979)	mAP50			77.40%	mAP50	74.80%	
87 imágenes (7,167 instancias)		mAP50-90			60.30%	mAP50-90	58.60%	
-			6	Sin registros	Precision	72.30%	Precision	77.40%
45°	22 (2,382)	Recall			55.50%	Recall	55.60%	
90°	15 (2,016)	mAP50			65.70%	mAP50	66.90%	
37 imágenes (4,398 instancias)		mAP50-90			44.40%	mAP50-90	47.70%	
-			7	Sin registros	Precision	81.40%	Precision	76.60%
45°	17 (2,071)	Recall			59.90%	Recall	58.60%	
90°	14 (2,181)	mAP50			72.40%	mAP50	68.70%	
31 imágenes (4,252 instancias)		mAP50-90			55.80%	mAP50-90	49.80%	
-			8	Sin registros	Precision	51.20%	Precision	63.20%
45°	10 (805)	Recall			34.20%	Recall	45.50%	
90°	10 (1,244)	mAP50			37.20%	mAP50	52.40%	
20 imágenes (2,049 instancias)		mAP50-90			20.10%	mAP50-90	33.70%	

Se observó que el mayor porcentaje de las predicciones se alcanza con el ángulo de captura a 90° y en las primeras tres semanas, ya que se reducen los errores de falsos positivos, puntuando un promedio de 82.83% de *Precision*. Sin embargo, las incidencias de falsos negativos son altas ya que, en promedio se registró un 72.13% lo que significa una exclusión de las hojas existentes en la imagen. Respecto a la forma del cuadro delimitador, se nota una mayor precisión de la región de las hojas puesto que, las métricas *mAP@50* y *mAP@50-95* alcanzaron en promedio 81.66% y 66.86% respectivamente.

A partir de la cuarta semana y hasta la octava, los indicadores mostraron puntajes mixtos. En la cuarta semana *Precision* y *mAP@50* son superiores al 80% lo que indica una correcta delimitación de las hojas detectadas. No obstante, la métrica *Recall* alcanza el 73.0% lo que representa una baja sensibilidad.

Los resultados de la quinta semana (concretamente en las imágenes capturadas a 45°) el porcentaje de *Precision* se mantuvo por encima del 80% señalando una correcta detección de las hojas, pero con un *Recall* bajo del 67.0%.

Entre la sexta y séptima semana, se observó un retroceso promedio de 57.40% en la métrica *Recall*. El porcentaje promedio del indicador *Precision* cae también a 76.92%; la clasificación se degrada progresivamente. Asimismo, la demarcación precisa de las hojas baja a niveles promedio de 68.42% *mAP@50* y 49.42% de *mAP@50-95* respectivamente. Finalmente, en la octava semana el desempeño es pobre en ambos casos de estudio (45° y 90°) dado que las cuatro métricas reportan valores cercanos al 50%.

En la Tabla 7.32 (pp. 101 - 102) se muestran 19 casos de evaluación con imágenes del cultivo de cacahuete, donde se emplea *YoloLeaf* para la localización de las hojas.

Tabla 7.32. Evaluación cualitativa del cacahuete.

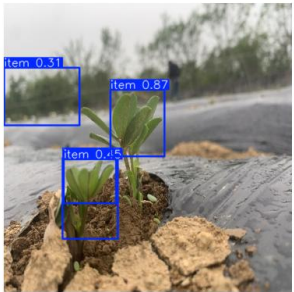
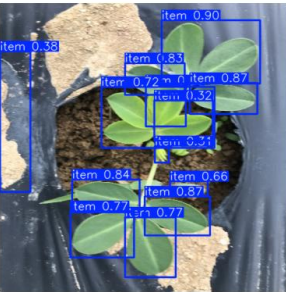
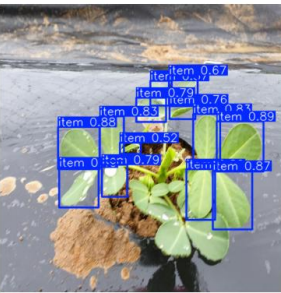
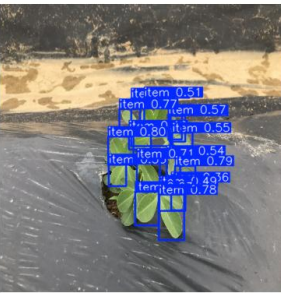
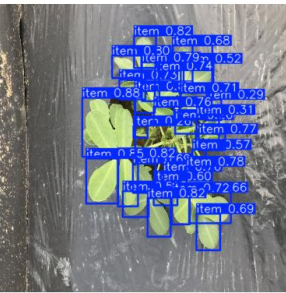
Semana	Ângulo 0°	Ângulo 45°	Ângulo 90°
1			
2			
3			

Tabla 7.32. Evaluación cualitativa del cacahuete.

Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
4	Sin registros		
5	Sin registros		
6	Sin registros		
7	Sin registros		
8	Sin registros		

En las primeras tres semanas se presentan problemas para hacer una correcta detección donde figuran las hojas en los ángulos de captura de 0° y 45°. Respecto al ángulo 0°, se registraron más falsos positivos debido a la presencia de follaje verde en el fondo de la imagen; la perspectiva no es favorable.

Desde la semana cuatro hasta la ocho, se complica realizar la inspección visual hoja por hoja del cultivo debido a las múltiples oclusiones que se originan a raíz del conglomerado de hojas existentes. De igual manera la distancia de captura es otro factor por considerar ya que, se reduce la escala de las hojas lo que dificulta la obtención de información para una correcta inferencia del modelo. Las imágenes de la semana seis y ocho con un ángulo de visión a 45° son evidencias donde se notó un pobre desempeño de *YoloLeaf* a causa de la perspectiva.

Las imágenes con el ángulo de 90° permitieron tener una visión favorable de la planta y con ello mejores resultados en la mayoría de las semanas. Sin embargo, en la semana uno se observó que el modelo presenta un falso positivo con una región de tierra y problemas en la localización precisa de las láminas foliares.

b) Frijol

El estudio se realizó con una adquisición de 645 imágenes que contienen un total de 10,220 cuadros delimitadores. En la Tabla 7.33 (pp. 103 - 104) se detallan 19 casos donde se cuantifica la inferencia de *YoloLeaf* acorde al número de semana y capturas.

Tabla 7.33. Evaluación cuantitativa del frijol.

Ángulo	Imágenes	Semana	Ángulo 0°		Ángulo 45°		Ángulo 90°	
0°	71 (290)	1	Precision	80.00%	Precision	82.50%	Precision	73.40%
45°	78 (378)		Recall	69.10%	Recall	71.70%	Recall	61.10%
90°	30 (175)		mAP50	77.50%	mAP50	80.60%	mAP50	68.10%
179 imágenes (843 instancias)			mAP50-90	64.00%	mAP50-90	73.00%	mAP50-90	59.80%
0°	28 (155)	2	Precision	76.00%	Precision	86.00%	Precision	90.10%
45°	78 (332)		Recall	72.90%	Recall	77.80%	Recall	81.60%
90°	38 (176)		mAP50	77.80%	mAP50	86.70%	mAP50	91.10%
144 imágenes (663 instancias)			mAP50-90	66.30%	mAP50-90	80.10%	mAP50-90	83.90%
0°	10 (94)	3	Precision	73.30%	Precision	82.60%	Precision	91.00%
45°	57 (616)		Recall	70.20%	Recall	73.20%	Recall	71.60%
90°	29 (298)		mAP50	78.20%	mAP50	84.80%	mAP50	85.50%
96 imágenes (1,008 instancias)			mAP50-90	64.80%	mAP50-90	70.70%	mAP50-90	76.30%
-		4	Sin registros		Precision	90.00%	Precision	91.10%
45°	84 (1,720)				Recall	82.00%	Recall	80.00%
90°	40 (1,189)				mAP50	91.50%	mAP50	91.20%
124 imágenes (2,909 instancias)					mAP50-90	81.10%	mAP50-90	81.60%
-		5	Sin registros		Precision	89.40%	Precision	88.40%
45°	28 (965)				Recall	73.20%	Recall	73.40%
90°	14 (481)				mAP50	86.70%	mAP50	86.70%
42 imágenes (1,446 instancias)					mAP50-90	74.30%	mAP50-90	71.60%
-		6	Sin registros		Precision	83.20%	Precision	88.40%
45°	10 (522)				Recall	71.80%	Recall	71.60%
90°	10 (524)				mAP50	81.90%	mAP50	85.50%
20 imágenes (1,046 instancias)					mAP50-90	66.00%	mAP50-90	69.50%

Ángulo	Imágenes	Semana	Ángulo 0°	Ángulo 45°		Ángulo 90°	
-	-	7	Sin registros	<i>Precision</i>	83.00%	<i>Precision</i>	90.30%
45°	10 (589)			<i>Recall</i>	67.20%	<i>Recall</i>	77.90%
90°	10 (493)			<i>mAP50</i>	79.50%	<i>mAP50</i>	89.00%
20 imágenes (1,082 instancias)				<i>mAP50-90</i>	63.60%	<i>mAP50-90</i>	78.60%
-	-	8	Sin registros	<i>Precision</i>	87.10%	<i>Precision</i>	78.40%
45°	10 (556)			<i>Recall</i>	63.30%	<i>Recall</i>	71.70%
90°	10 (667)			<i>mAP50</i>	75.20%	<i>mAP50</i>	80.10%
20 imágenes (1,223 instancias)				<i>mAP50-90</i>	60.80%	<i>mAP50-90</i>	65.70%

De acuerdo con la información registrada se observaron resultados mixtos en las cuatro métricas de evaluación durante las ocho semanas en los ángulos de encuadre de 45° y 90°.

Se identificó que en la segunda semana se anotaron 90.10% de *Precision*, 81.10% de *Recall* minimizando los errores de falsos positivos y negativos en 78 imágenes que alojan 378 instancias. De igual modo, *mAP@50* y *mAP@50-95* lograron 91.10% y 83.90% respectivamente, reduciendo las impresiones de ubicación de las hojas en imágenes con un campo de visión de 90°.

Además, se observó que los cuatro indicadores arrojaron puntajes altos de 90.30% *Precision*, 77.90% de *Recall*, 89.0% de *mAP@50* y 78.60% de *mAP@50-95* en imágenes con un encuadre a 90° en la séptima semana, reduciendo así los errores de inferencia tanto en la clasificación (casos falsos positivos y negativos) como en la estimación de las regiones de interés que abarcan las hojas.

En la Tabla 7.34 (pp. 104 - 106) se muestran 19 ejemplares capturadas a campo abierto que ilustran las detecciones de hojas de la planta del frijol, utilizando el modelo *YoloLeaf*.

Tabla 7.34. Evaluación cualitativa del frijol.

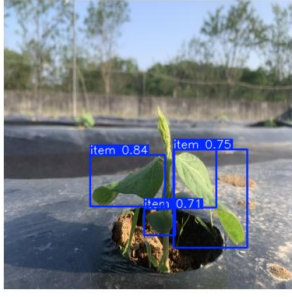
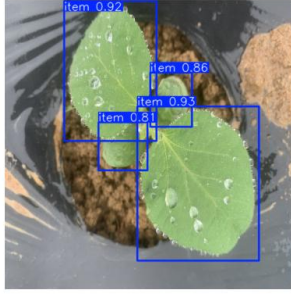
Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
1			
2			

Tabla 7.34. Evaluación cualitativa del frijol.

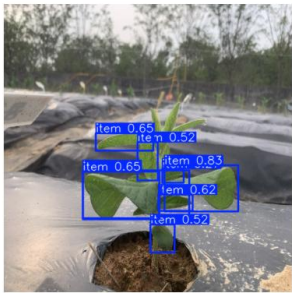
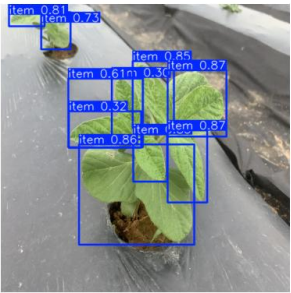
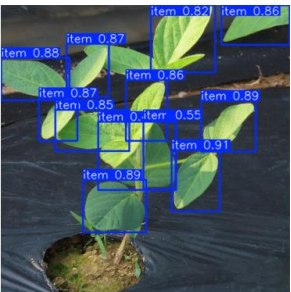
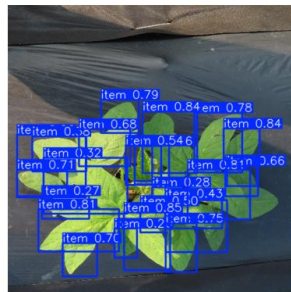
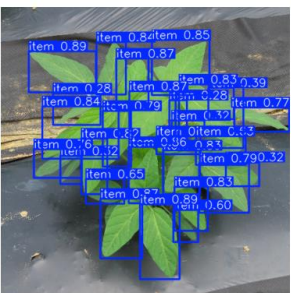
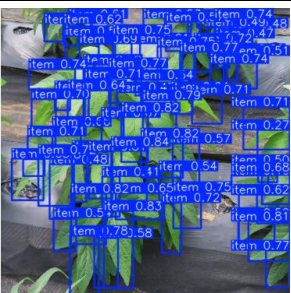
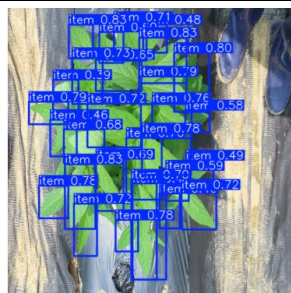
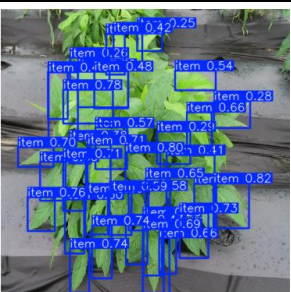
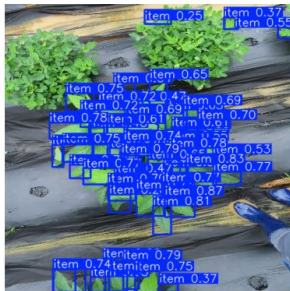
Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
3			
4	Sin registros		
5	Sin registros		
6	Sin registros		
7	Sin registros		

Tabla 7.34. Evaluación cualitativa del frijol.

Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
8	Sin registros		

En las primeras cuatro semanas de crecimiento de la planta se apreciaron algunos errores de inferencia del modelo propuesto. Por ejemplo, en las imágenes cuyos encuadres son de 0° y 90° en la primera semana, se notaron problemas en detectar las hojas que figuran en la imagen (falsos negativos) afectando negativamente la métrica *Recall*. Sin embargo, el ángulo de toma a 45° resultó ser favorable en esta semana al detectar correctamente las láminas foliares.

Para el caso de la segunda semana, la perspectiva de captura a 90° fue positiva ya que, se muestra un correcto desempeño de *YoloLeaf* al detectar todas las hojas de la planta que presentan gotas de agua en la superficie foliar. No obstante, se observó una superposición de *BBs* debido a la forma de la hoja que adopta en la muestra con perspectiva a 0°.

En la cuarta semana es donde se registró el porcentaje de *Recall* más alto (ver Tabla 7.33), *YoloLeaf* logró detectar correctamente la hoja ocluida por el tallo y ramas del cultivo en la imagen tomada a 45°. Sin embargo, a 90° se nota el ápice de una hoja sin ser detectada a causa del solapamiento entre hojas.

De la semana quinta hasta la octava, es complicado dar un seguimiento cualitativo hoja por hoja. No obstante, la evaluación cualitativa mostró un buen desempeño a simple vista en la semana siete en imágenes capturadas a 90°. El análisis cuantitativo en la Tabla 7.33 justifica el funcionamiento de *YoloLeaf* en este caso de estudio.

c) Frijol rojo

El análisis se llevó a cabo con una selección de 663 imágenes que alojan un total de 8,597 instancias. En la Tabla 7.35 (*pp.* 106 - 107) se detallan 20 casos donde se mide la inferencia de *YoloLeaf* acorde al número de semana y captura.

Tabla 7.35. Evaluación cuantitativa del frijol rojo.

Ángulo	Imágenes	Semana	Ángulo 0°		Ángulo 45°		Ángulo 90°	
0°	63 (150)	1	Precision	87.80%	Precision	78.50%	Precision	87.10%
45°	71 (223)		Recall	81.90%	Recall	72.60%	Recall	82.50%
90°	37 (154)		mAP50	87.30%	mAP50	79.50%	mAP50	89.00%
171 imágenes (527 instancias)			mAP50-90	71.10%	mAP50-90	65.40%	mAP50-90	86.10%
0°	22 (98)	2	Precision	86.70%	Precision	79.80%	Precision	92.40%
45°	110 (399)		Recall	78.60%	Recall	78.10%	Recall	80.50%
90°	45 (215)		mAP50	84.20%	mAP50	84.70%	mAP50	91.90%
177 imágenes (712 instancias)			mAP50-90	63.10%	mAP50-90	72.30%	mAP50-90	84.60%
0°	10 (60)		Precision	90.10%	Precision	85.90%	Precision	85.00%

Ángulo	Imágenes	Semana	Ángulo 0°		Ángulo 45°		Ángulo 90°	
45°	54 (501)	3	Recall	73.30%	Recall	82.20%	Recall	82.20%
90°	28 (235)		mAP50	82.20%	mAP50	89.30%	mAP50	89.00%
92 imágenes (796 instancias)			mAP50-90	72.00%	mAP50-90	78.70%	mAP50-90	79.90%
0°	10 (130)	4	Precision	75.90%	Precision	89.70%	Precision	92.00%
45°	57 (1,031)		Recall	72.60%	Recall	87.10%	Recall	88.70%
28 / 433 / 90°			mAP50	79.80%	mAP50	93.40%	mAP50	93.70%
95 imágenes (1,594 instancias)			mAP50-90	62.80%	mAP50-90	85.20%	mAP50-90	86.70%
-		5	Sin registros		Precision	84.50%	Precision	88.50%
45°	41 (1,289)				Recall	69.60%	Recall	72.90%
90°	27 (1,019)				mAP50	82.70%	mAP50	85.70%
68 imágenes (2,308 instancias)					mAP50-90	67.90%	mAP50-90	73.60%
-		6	Sin registros		Precision	81.60%	Precision	85.40%
45°	10 (439)				Recall	71.30%	Recall	63.10%
90°	10 (500)				mAP50	82.90%	mAP50	77.60%
20 imágenes (939 instancias)					mAP50-90	67.70%	mAP50-90	61.70%
-		7	Sin registros		Precision	73.60%	Precision	83.80%
45°	10 (520)				Recall	64.60%	Recall	62.00%
90°	10 (497)				mAP50	75.80%	mAP50	78.10%
20 imágenes (1,017 instancias)					mAP50-90	58.80%	mAP50-90	62.30%
-		8	Sin registros		Precision	83.80%	Precision	88.90%
45°	10 (461)				Recall	74.20%	Recall	78.70%
90°	10 (503)				mAP50	85.10%	mAP50	89.80%
20 imágenes (964 instancias)					mAP50-90	73.40%	mAP50-90	81.30%

Se observó que los ejemplares capturados con un ángulo a 90° resultó ser favorable en la mayoría de las semanas (excepto semana tres y seis) dado que, es donde se registraron los porcentajes más altos en las métricas de evaluación.

La tasa menor de falsos positivos ocurrió en la semana dos y cuatro al puntuar 92.40% y 92.0% de *Precision* respectivamente. Mientras que, los casos de falsos negativos fueron menores en las semanas uno y cuatro al alcanzar 82.50% y 88.70% de *Recall* respectivamente. Con respecto a la cuantificación de las instancias detectadas, los mayores puntajes se reportaron en las semanas dos y cuatro logrando 91.90% de *mAP@50*, 84.60% de *mAP@50-95* y 93.70% de *mAP@50*, 86.70% de *mAP@50-95* respectivamente. En conclusión, la semana cuatro fue la mejor opción para detectar correctamente las hojas al evaluar 433 instancias en 28 imágenes a un ángulo de visión a 90°.

De la semana cinco hasta la octava se notó que el porcentaje promedio de la métrica *Precision* fue de 86.65%, mostrando el puntaje más alto en la semana ocho de 88.90%. Asimismo, *mAP@50* y *mAP@50-95* alcanzaron los dígitos más altos de 89.80% y 81.30% respectivamente. Sin embargo, el valor porcentual promedio del indicador *Recall* fue de 69.17% debido a que en las semanas seis y siete se registraron 63.10% y 62.0% respectivamente. En resumen, se cuantificó una precisión aceptable que reduce los casos falso positivos. No obstante, la sensibilidad baja origina una alta probabilidad de falsos negativos.

En la Tabla 7.36 (pp. 108 - 109) se muestran 20 ejemplos capturados en diferentes condiciones al exterior, donde figuran las predicciones hechas de *YoloLeaf* con hojas de la planta del frijol rojo.

Tabla 7.36. Evaluación cualitativa del frijol rojo.

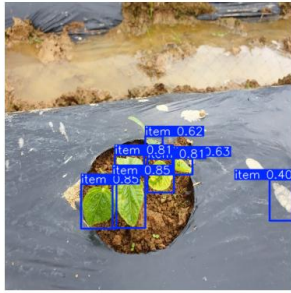
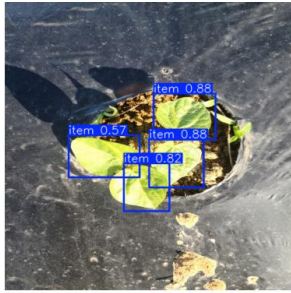
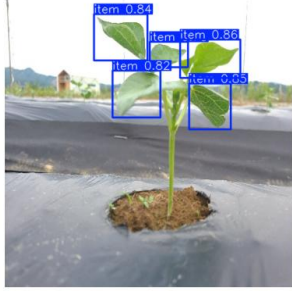
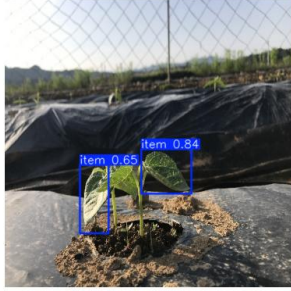
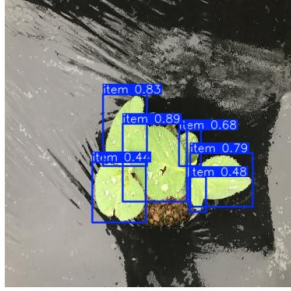
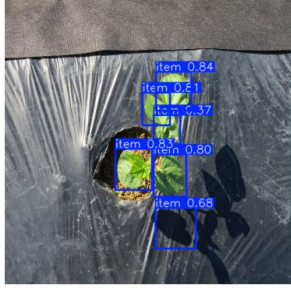
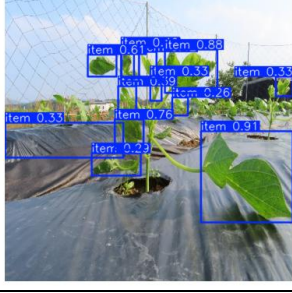
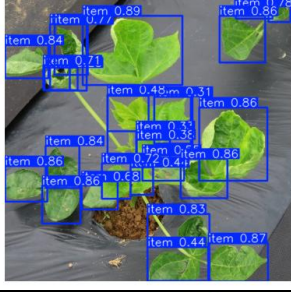
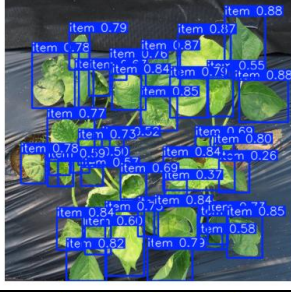
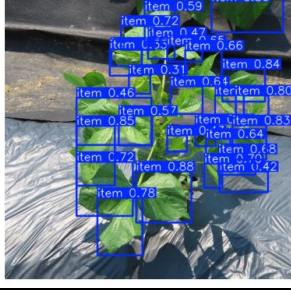
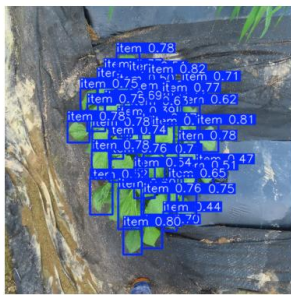
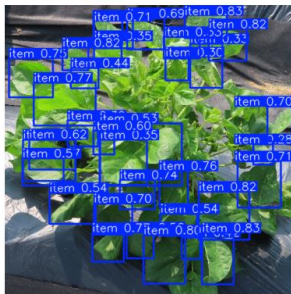
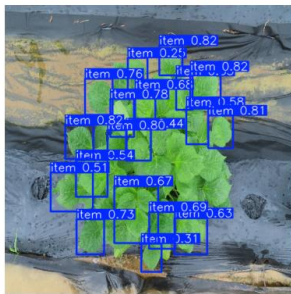
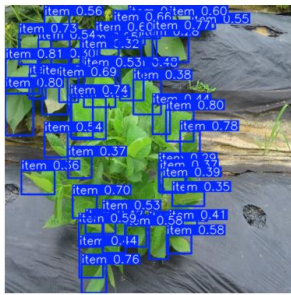
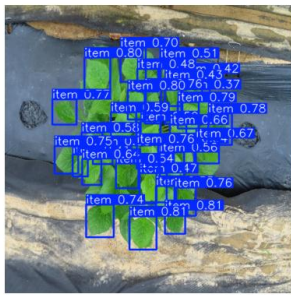
Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
1			
2			
3			
4			
5	Sin registros		

Tabla 7.36. Evaluación cualitativa del frijol rojo.

Semana	Ángulo 0°	Ángulo 45°	Ángulo 90°
6	Sin registros		
7	Sin registros		
8	Sin registros		

En la semana uno se apreciaron errores de inferencias del tipo falsos positivos en las tomas a 0° y 45° ya que, la presencia de tierra y manchas son localizados y etiquetados como hojas. Estas fallas se cuantifican con la métrica *Precision*. También, se presentaron imprecisiones para demarcar correctamente la región donde figura la lámina foliar tal y como se ilustra en la captura a 90° afectando negativamente los indicadores *mAP@50* y *mAP@50-95*.

En la semana dos y tres se observaron problemas de predicción del tipo falsos negativos en las perspectivas 0° y 45° puesto que, algunas de las hojas presentes no son detectados por el modelo. Estos desaciertos se miden con el indicador *Recall*. Cabe mencionar que en la misma semana y la semana cuatro, se presentaron estimaciones erróneas de hojas que presentaron daños o anomalías su forma en captura a 45°.

A partir de la semana cinco hasta la ocho, la evaluación cualitativa se vuelve un reto para dar el seguimiento hoja por hoja. No obstante, la evaluación cuantitativa mostrada en la Tabla 7.35 apunta a un desempeño aceptable en cada caso señalado (excepto en tomas a 45° y 90° en la semana siete). En la semana ocho se observaron múltiples detecciones pese a la escala de las hojas y el encuadre a 90° del cultivo.

7.10 Comparación cuantitativa: Yolov12n, Yolov11n vs YoloLeaf

En la Tabla 7.37 se sintetizan únicamente los datos que corresponden a las imágenes cuya perspectiva de captura es a 90° (ángulo favorable) de las semanas uno, cuatro y ocho.

Tabla 7.37. Síntesis comparativa de Yolov12n, Yolov11n y YoloLeaf.

Semana 1						
Modelo	Cacahuete		Frijol		Frijol rojo	
Yolov12n	Precision	67.70%	Precision	61.40%	Precision	85.10%
	Recall	74.40%	Recall	59.40%	Recall	68.20%
	mAP50	73.30%	mAP50	62.60%	mAP50	80.90%
	mAP50-90	49.50%	mAP50-90	48.00%	mAP50-90	69.20%
Yolov11n	Precision	78.70%	Precision	69.10%	Precision	85.00%
	Recall	74.20%	Recall	49.70%	Recall	64.30%
	mAP50	81.40%	mAP50	57.50%	mAP50	80.00%
	mAP50-90	58.10%	mAP50-90	45.50%	mAP50-90	69.50%
YoloLeaf	Precision	81.60%	Precision	73.40%	Precision	87.10%
	Recall	76.00%	Recall	61.10%	Recall	82.50%
	mAP50	83.80%	mAP50	68.10%	mAP50	89.00%
	mAP50-90	66.50%	mAP50-90	59.80%	mAP50-90	86.10%
Semana 4						
Yolov12n	Precision	82.50%	Precision	88.10%	Precision	87.00%
	Recall	71.70%	Recall	81.20%	Recall	83.10%
	mAP50	80.40%	mAP50	91.50%	mAP50	91.40%
	mAP50-90	56.00%	mAP50-90	72.10%	mAP50-90	73.50%
Yolov11n	Precision	81.20%	Precision	89.10%	Precision	86.00%
	Recall	71.00%	Recall	80.10%	Recall	83.70%
	mAP50	80.80%	mAP50	91.00%	mAP50	91.40%
	mAP50-90	58.20%	mAP50-90	74.90%	mAP50-90	75.40%
YoloLeaf	Precision	82.20%	Precision	91.10%	Precision	92.00%
	Recall	73.00%	Recall	80.00%	Recall	88.70%
	mAP50	81.30%	mAP50	91.20%	mAP50	93.70%
	mAP50-90	66.50%	mAP50-90	81.60%	mAP50-90	86.70%
Semana 8						
Yolov12n	Precision	47.70%	Precision	75.80%	Precision	79.50%
	Recall	41.00%	Recall	65.90%	Recall	78.50%
	mAP50	37.90%	mAP50	75.90%	mAP50	86.80%
	mAP50-90	16.20%	mAP50-90	53.50%	mAP50-90	64.20%
Yolov11n	Precision	48.10%	Precision	77.30%	Precision	80.10%
	Recall	38.20%	Recall	68.50%	Recall	73.60%
	mAP50	39.30%	mAP50	79.20%	mAP50	84.40%
	mAP50-90	18.40%	mAP50-90	57.20%	mAP50-90	64.40%
YoloLeaf	Precision	63.20%	Precision	78.40%	Precision	88.90%
	Recall	45.50%	Recall	71.70%	Recall	78.70%
	mAP50	52.40%	mAP50	80.10%	mAP50	89.80%
	mAP50-90	33.70%	mAP50-90	65.70%	mAP50-90	81.30%

Las métricas señalan que YoloLeaf demuestra un buen desempeño al reducir los errores de inferencia en la semana uno y ocho en los tres cultivos estudiados.

Por ejemplo, en las imágenes de los cultivos de cacahuete y frijol existen cambios porcentuales considerable en la semana uno con respecto a Yolov12n y YoloLeaf de 20.53% y 19.54% de Precision respectivamente, lo que señala una probabilidad menor de falsos positivos. Asimismo, entre los modelos Yolov12n y YoloLeaf se cuantificó un cambio porcentual de 32.49% con respecto a la métrica Precision en la semana ocho, pero sólo en imágenes del cultivo de cacahuete.

El siguiente dato que resaltar ocurre con capturas del frijol rojo en la semana uno entre los modelos *Yolov12n* y *YoloLeaf*, registrando un cambio porcentual del 20.96% con respecto al indicador de *Recall*, lo que significa que el modelo propuesto es menos susceptible en casos de falsos negativos.

Existe una reducción importante de falsos negativos entre *Yolov11n* vs *YoloLeaf*, un cambio porcentual del 28.30% referente al índice *Recall* en capturas donde figura la planta del frijol rojo de la semana uno.

En la semana cuatro los cultivos de cacahuate y frijol se reportan resultados mixtos entre *Yolov12n* y *YoloLeaf*. El modelo propuesto logró un porcentaje superior de 81.60% de *mAP50-90* contra el 72.10% de *mAP50-90* generado por *Yolov12n*, un cambio porcentual de 13.17%, demostrando que *YoloLeaf* es más preciso en hacer una correcta delimitación de la región donde figuran las hojas del frijol.

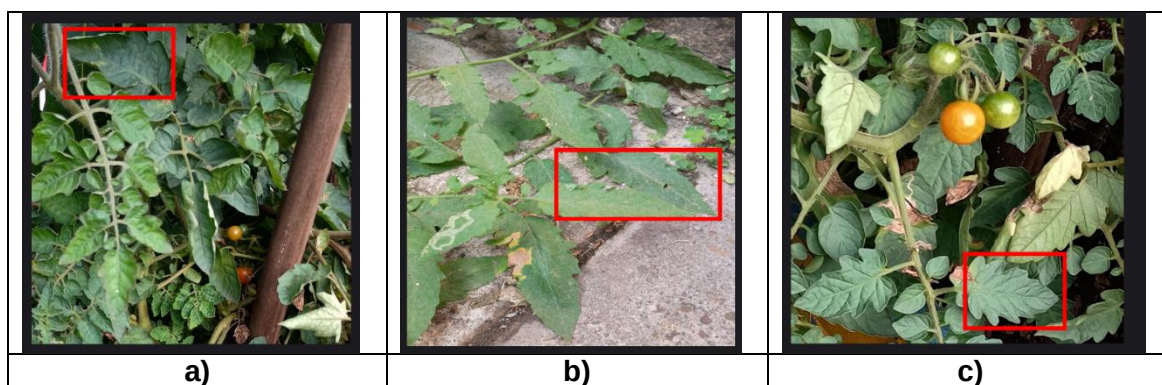
En la semana ocho los mejores resultados se registran con el modelo *YoloLeaf* en capturas de la planta del frijol rojo, donde se alcanza una menor probabilidad de falsos positivos contra *Yolov12n*, lo que significa un cambio porcentual de 11.82% en relación con *Precision*. Adicionalmente la demarcación de los cuadros limitadores es más precisa que *Yolov12n*, al reportarse un cambio porcentual del 26.63% en relación al índice *mAP50-90*.

Los tres modelos presentaron un desempeño pobre en las imágenes de los cultivos del cacahuate en la semana ocho. Pese a los resultados bajos, *YoloLeaf* demuestra ser mejor contra *Yolov11n* y *Yolov12n* en estas circunstancias adversas.

a) Resultados de localización en la aplicación móvil

Los tres ejemplares contemplados para este análisis fueron tomados con un teléfono celular (*Motorola G55 5G*) cuya resolución es de 1,024×1,024 píxeles. En la Tabla 7.38 se presentan tres capturas donde se visualiza la localización de una hoja, empleando el modelo desarrollado *YoloLeaf*.

Tabla 7.38. Evaluación cualitativa.



En a) se aprecia una correcta delimitación del espacio donde yace la hoja de la planta. Además, dentro de la región de interés existe una pequeña lámina. En b) a simple vista pareciera una falla, sin embargo, la localización se centra en la hoja que presenta oclusiones parciales a causa del follaje. En c) se nota una mínima imprecisión, pero se detecta la mayor región donde está la lámina foliar.

La Tabla 7.39 detalla la evaluación del tiempo de inferencia y memoria utilizados, seleccionado 10 ejemplares tomados al azar, preservando siempre una correcta localización del objetivo tal y como se ilustra en la Tabla 7.38, caso a). En este análisis se resaltan en color azul las cifras menores (mejores resultados).

Tabla 7.39. Evaluación cuantitativa del desempeño.

a)		b)		c)	
Tiempo (ms)	Memoria (Mbytes)	Tiempo (ms)	Memoria (Mbytes)	Tiempo (ms)	Memoria (Mbytes)
178	7	179	5	176	5
199	10	176	10	175	21
177	12	175	20	178	23
176	15	182	22	206	28
176	17	179	9	180	7
176	19	176	27	178	16
177	22	206	23	174	21
177	24	174	18	187	23
181	19	200	26	210	19
175	14	196	22	199	10
Promedio		Promedio		Promedio	
179.20	15.90	184.30	18.20	186.30	17.30

De acuerdo con la Tabla 7.39 los puntajes menores del tiempo de inferencia y memoria son 174 ms y 5 Mbytes respectivamente. Sin embargo, a) logra en promedio obtener los menores puntajes tanto en tiempo de inferencia (179.20 ms) y uso de memoria RAM (15.90 Mbytes) en el dispositivo móvil respectivamente.

7.11 Resultados de la segmentación de hojas

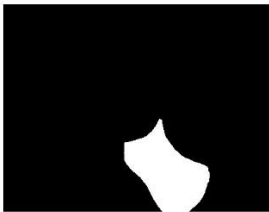
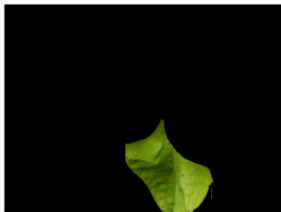
Una vez localizadas las hojas por medio de los BBs generados por *YoloLeaf*, estas se emplean como un *prompt* que recibe el modelo *Sam2Tiny* para inferir el área de píxeles que se aproximan a la forma de la lámina foliar real.

Las Tablas 7.40 - 7.52 cuantifican la semejanza de la máscara de píxel predicha con respecto al objetivo bajo análisis.

- **Oclusiones**

En la siguiente imagen (ver Tabla 7.40) existen seis hojas de la planta del jitomate. El ejemplar seleccionado muestra una doble oclusión por dos hojas del tipo papa que presentan una serie de anomalías (cambios de coloración y manchas de color marrón) en la superficie.

Tabla 7.40. Análisis de resultados de segmentación (oclusión).

Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.89%
			F1	99.66%
			IoU	99.33%
			Precision	99.45%
			Recall	99.87%

La predicción que realizó *Sam2Tiny* fue precisa ya que, la máscara de píxeles corresponde a la forma de la hoja protagonista, excluyendo las otras dos regiones de láminas foliares. Los puntajes marcados en los cinco indicadores confirman una alta precisión del aspecto de la máscara predicha con respecto al *GT*.

- **Cambios en la forma**

La siguiente captura (ver Tabla 7.41) se ejemplifica un cambio anómalo significativo en la forma regular de los tres ejemplares debido a que las hojas adquieren un doblez pronunciado, dejando expuesto de manera parcial el envés y cubriendo casi en su totalidad el área principal de la lámina foliar. La hoja elegida muestra de manera incompleta ambas partes de la hoja (haz y envés).

Tabla 7.41. Análisis de resultados de segmentación (cambio de forma).

Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.53%
			F1	98.79%
			IoU	97.61%
			Precision	99.99%
			Recall	97.62%

Aun cuando la alteración en la hoja es notoria, la predicción efectuada por *Sam2Tiny* mostró un buen desempeño en generar la máscara de píxeles equivalente al *GT*. Los puntajes que reflejan las imprecisiones minúsculas son *IoU* y *Recall* ya que, *IoU* mide el área en común entre la máscara predicha y la máscara del *GT* y, *Recall* calcula los píxeles que pertenecen a la hoja y no al fondo de la imagen.

- **Daño por plagas**

La imagen de la Tabla 7.42 ilustra uno de los innumerables retos de la segmentación de imágenes puesto que, el objetivo muestra cambios importantes en la textura de la lámina principal. Adicionalmente, las perforaciones sobre el área de interés suman complejidad ya que, si por las rupturas figura fondo idéntico al de la hoja podría no segmentar correctamente la región donde no existe parte de la hoja.

Tabla 7.42. Análisis de resultados de segmentación (daño por plaga).

Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.78%
			F1	99.46%
			IoU	98.94%
			Precision	99.96%
			Recall	98.97%

Las métricas indican que *Sam2Tiny* en este ejemplar no se presentaron problemas mayores, puesto que se logró delimitar correctamente el área pese a las condiciones físicas de la hoja, generando una máscara de píxeles idéntica al *GT*.

- **Agrupacion de hojas no anomalas**

En la Tabla 7.43 se ilustra una agrupación de hojas sanas del tipo sierra de la planta del jitomate. Sin embargo, el objetivo detectado en este caso muestra un cambio importante en la forma de la lámina foliar, dejando expuesto de manera parcial el envés.

Tabla 7.43. Análisis de resultados de segmentación (cambio en aspecto).

Entrada	<i>Ground Truth</i>	Salida	Métricas	
			<i>Accuracy</i>	99.56%
			<i>F1</i>	98.49%
			<i>IoU</i>	97.03%
			<i>Precision</i>	99.23%
			<i>Recall</i>	97.77%

Tomando como referencia los indicadores y la inspección visual de la máscara predicha por *Sam2Tiny* se observa una segmentación con ligeros errores puesto que, la minúscula área oscura que se produce debido al dobles de la hoja no fue segmentado por *SAM2*. Las métricas que corroboran esta incidencia son *IoU* y *Recall* registrando un 97.03% y 97.77% respectivamente.

- **Iluminación irregular**

El siguiente ejemplo (ver Tabla 7.44) muestra una condición adversa a causa de la iluminación irregular, lo que conlleva a la proyección de una serie de sombras sobre las hojas aparentemente en condiciones saludables de la planta del jitomate. La mayor área parcial del objeto de interés es cubierta por una sombra.

Tabla 7.44. Análisis de resultados de segmentación (iluminación irregular).

Entrada	<i>Ground Truth</i>	Salida	Métricas	
			<i>Accuracy</i>	99.85%
			<i>F1</i>	99.46%
			<i>IoU</i>	98.93%
			<i>Precision</i>	99.75%
			<i>Recall</i>	99.17%

Pese al cambio de coloración a causa de la iluminación irregular sobre la superficie parcial de la lámina foliar, el modelo *Sam2Tiny* logró generar la máscara de píxeles precisa con minúsculos errores que se ven cuantificados en 98.93% de *IoU* y 99.17% de *Recall*.

- **Presencia de jitomates I**

La siguiente captura (ver Tabla 7.45) es realizada hacia una agrupación de múltiples hojas en condiciones saludables del tipo papa de la planta de jitomate, que presentan diferentes

dimensiones y oclusiones. El objeto de interés seleccionado se ubica en una región donde otras láminas foliares están solapándolo de manera parcial.

Tabla 7.45. Análisis de resultados de segmentación (solapamientos).

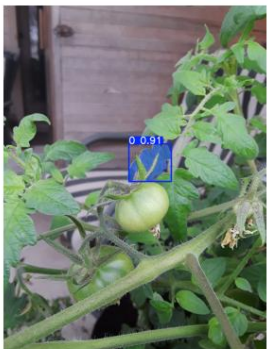
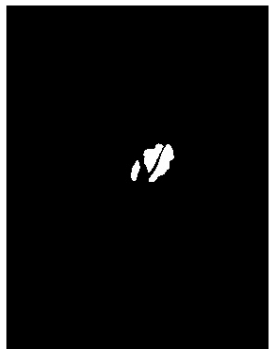
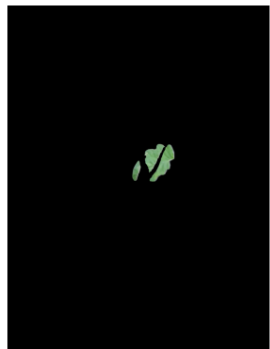
Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.97%
			F1	99.84%
			IoU	99.69%
			Precision	99.99%
			Recall	99.69%

La región de interés definida por el cuadro delimitador involucra de manera parcial dos áreas de hojas que solapan al objeto de interés. En este caso la inferencia de la máscara de píxeles hecha por *Sam2Tiny* mostró un buen desempeño al excluir las otras láminas, ya que comparten características idénticas (coloración). Las métricas cercanas al 100% confirman el correcto desempeño en este caso.

- **Presencia de jitomates II**

En este caso (ver Tabla 7.46) se selecciona una hoja pequeña del tipo sierra que presenta una oclusión parcial a causa del sépalo del jitomate; ambos objetos comparten una tonalidad de color verde.

Tabla 7.46. Análisis de resultados de segmentación (oclusión).

Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.93%
			F1	99.72%
			IoU	99.44%
			Precision	99.99%
			Recall	99.44%

La inferencia de *Sam2Tiny* pudo delimitar correctamente el área que corresponde a la hoja, excluyendo en su totalidad el sépalo; las métricas de evaluación confirman el desempeño de la segmentación precisa.

- **Agrupación de hojas anómalas**

La Tabla 7.47 presenta una agrupación de hojas pertenecientes a la planta del jitomate en condiciones no saludables del tipo papa que manifiestan algún tipo de anomalía. El objeto detectado muestra una alteración tanto en la coloración (tonos amarillos, verdes y manchas marrones) como en la forma (aspecto particular en el perímetro). Además, dentro de la región de interés figura parcialmente la lámina de otra hoja con características similares en coloración.

Tabla 7.47. Análisis de resultados de segmentación (cambio en la tonalidad).

Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.89%
			F1	99.63%
			IoU	99.26%
			Precision	99.99%
			Recall	99.26%

La máscara creada por *Sam2Tiny* demostró una segmentación exacta de la hoja protagonista ya que, se excluye la otra lámina parcial. Los indicadores registrados señalan una predicción precisa; los indicadores están cercanos al 100%.

- **Iluminación irregular**

En la Tabla 7.48 se expone una situación compleja debido principalmente a la iluminación irregular y las múltiples orientaciones que presentan las hojas del tipo papa. El objeto de interés localizado se ubica en una región donde existen varias hojas. Asimismo, la superficie principal se encuentra parcialmente ocluida por el tipo de toma.

Tabla 7.48. Análisis de resultados de segmentación (proyección de sombras).

Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.88%
			F1	99.59%
			IoU	99.20%
			Precision	99.76%
			Recall	99.43%

La inferencia de *Sam2Tiny* creó una máscara de píxeles con diminutas imprecisiones en el ápice y base de la hoja. Sin embargo, los cambios de tonalidad en la superficie principal de la lámina no representaron problema, pues los indicadores arrojaron puntajes por encima del 99% en este caso de segmentación.

- **Presencia de insecto I**

La Tabla 7.49 ejemplifica un escenario donde la iluminación al proyectarse contra el objetivo de interés se genera una sombra sobre la lámina principal de la hoja, ocasionando cambios

en la coloración original. Además, dentro de la región que delimita el objeto figuran una parte del tallo y una minúscula parte de otra hoja.

Tabla 7.49. Análisis de resultados de segmentación (iluminación a contraluz).

Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.58%
			F1	98.61%
			IoU	97.27%
			Precision	99.38%
			Recall	97.86%

Basándose en los indicadores, se observa que la inferencia de *Sam2Tiny* presentó algunos fallos minúsculos ya que, se observa que la máscara predicha no segmentó la diminuta parte de la otra hoja que figura detrás de la lámina principal (cercas del ápice). Además, en la base de la lámina foliar no se logró recuperar completamente la extremidad derecha. Las incidencias recaen en *IoU* y *Recall* principalmente puntuando 97.27% y 97.86% respectivamente.

- **Presencia de insecto II**

La imagen de la Tabla 7.50 muestra otra escena compleja debido a la iluminación que incide en las hojas, lo que genera diferentes cambios de tonalidad de color verde. Asimismo, el fondo de la imagen y el gusano ambos de tonalidad verde podrían generarse errores en la delimitación de las láminas foliares.

Tabla 7.50. Análisis de resultados de segmentación (cambios de coloración).

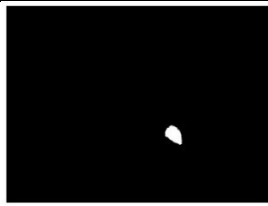
Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.92%
			F1	99.74%
			IoU	99.50%
			Precision	99.95%
			Recall	99.54%

Las métricas demuestran que *Sam2Tiny* superó los retos de este entorno ya que, la máscara de píxeles predicha se asemeja al *GT*, creando así una imagen segmentada precisa de la superficie de la hoja; sólo se presentó una minúscula inexactitud fuera del área de interés.

- **Presencia de objetos verdes**

En este ejemplo (ver Tabla 7.51) se expone una toma hacia una serie de macetas que contienen plantas de jitomate en etapa de desarrollo. El objetivo seleccionado en esta ocasión corresponde a una planta en etapa de crecimiento que está cubierta por un recipiente de plástico.

Tabla 7.51. Análisis de resultados de segmentación (hoja a través de objeto transparente).

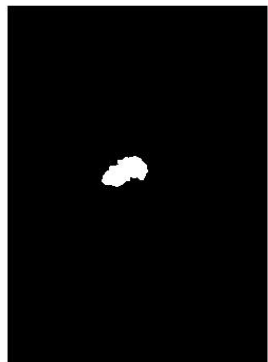
Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.98%
			F1	99.93%
			IoU	99.86%
			Precision	99.97%
			Recall	99.88%

Sam2Tiny logró crear una máscara de píxeles precisa, semejante al *GT* sin mayores percances, aun cuando la hoja se encuentra dentro de un recipiente de plástico transparente. Los valores de las métricas de evaluación confirmaron un buen desempeño con este ejemplar de escala pequeña.

- **Dibujos o impresiones**

En la Tabla 7.52 figura la captura una planta de jitomate en etapa de crecimiento. Las hojas de diferentes tamaños que figuran corresponden al tipo sierra aparentemente en condiciones saludables. El objetivo señalado se ubica en una región donde, existen otras hojas contiguas.

Tabla 7.52. Análisis de resultados de segmentación (lámina foliar no anómala).

Entrada	Ground Truth	Salida	Métricas	
			Accuracy	99.94%
			F1	99.79%
			IoU	99.58%
			Precision	99.99%
			Recall	99.59%

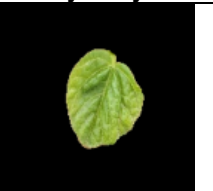
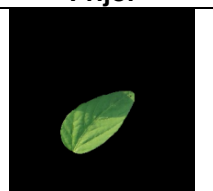
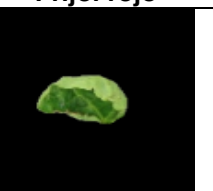
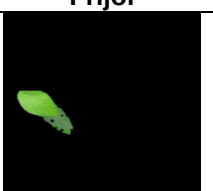
Las métricas de evaluación cuantificaron valores cercanos al 100%, por lo tanto, se determina que la inferencia de *Sam2Tiny* fue precisa al aproximarse correctamente al área definida en el *GT*.

7.12 Segmentación de hojas con otros tipos de cultivos

A continuación, se presenta la delimitación de hojas a partir de los *bounding boxes* generados con el modelo *YoloLeaf*. Como se mencionó se proporciona un *prompt* mejorado que guía a efectuar la segmentación con el modelo *Sam2Tiny*.

En la Tabla 7.53 se ilustran y detallan las segmentaciones de las láminas foliares de los cultivos de cacahuate, frijol y frijol rojo. Se seleccionan los ejemplares que corresponden a capturas realizadas a 90° durante las semanas uno, cuatro y ocho.

Tabla 7.53. Láminas foliares segmentadas.

Semana 1					
Cacahuate		Frijol		Frijol rojo	
					
Accuracy	99.95%	Accuracy	99.94%	Accuracy	99.94%
F1	99.80%	F1	99.75%	F1	99.77%
IoU	99.60%	IoU	99.50%	IoU	99.55%
Precision	99.90%	Precision	99.90%	Precision	99.99%
Recall	99.60%	Recall	99.50%	Recall	99.55%
Semana 4					
Cacahuate		Frijol		Frijol rojo	
					
Accuracy	99.95%	Accuracy	99.95%	Accuracy	99.96%
F1	99.79%	F1	99.79%	F1	99.83%
IoU	99.59%	IoU	99.59%	IoU	99.67%
Precision	99.90%	Precision	99.99%	Precision	99.90%
Recall	99.59%	Recall	99.59%	Recall	99.67%
Semana 8					
Cacahuate		Frijol		Frijol rojo	
					
Accuracy	98.86%	Accuracy	99.79%	Accuracy	99.94%
F1	98.31%	F1	98.96%	F1	99.76%
IoU	97.63%	IoU	97.95%	IoU	99.52%
Precision	98.72%	Precision	98.06%	Precision	99.54%
Recall	98.90%	Recall	99.88%	Recall	99.81%

La segmentación de imágenes en la primera y cuarta semana se obtienen buenos resultados ya que, no existen oclusiones de manera pronunciada entre las hojas. Se observó que *Sam2Tiny* logró inferir de manera correcta diferentes formas y tamaños de las láminas. Sin embargo, la complejidad crece a medida que los cultivos continúan en desarrollo.

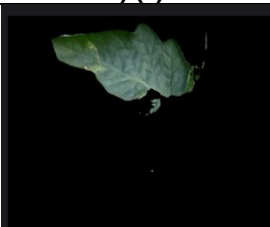
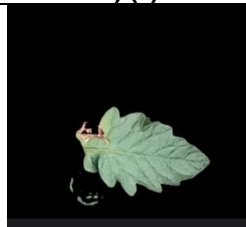
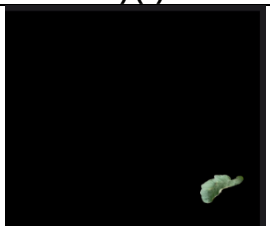
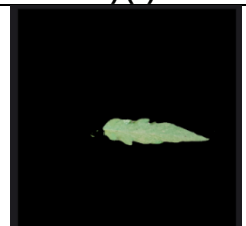
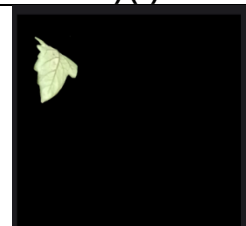
Al cabo de las ocho semanas la segmentación de elemento por elemento se vuelve un desafío debido a los solapamientos que se da entre las láminas foliares. Además, si otras hojas figuran en la región acotada por el cuadro delimitador como fondo, es posible la aparición de pequeñas partículas en la imagen final, debido a que la lámina foliar principal comparte determinadas características (mismo cultivo).

En esta última experimentación se observó la inferencia genérica (*Zero-Shot*) del modelo *Sam2Tiny* que lo distingue de otros modelos de segmentación ya que, por medio del *prompt* generado por *YoloLeaf* se efectuaron múltiples delimitaciones de las áreas correspondientes a las hojas en cada cultivo sin presentar problemas mayores.

b) Resultados de segmentación en la aplicación móvil

En la Tabla 7.54 se ejemplifican seis segmentaciones predichas con el modelo *MobileSAM*, donde sólo figura una lámina foliar por imagen que corresponden a los casos estudiados: a), b) y c) de la Tabla 7.38. El dígito que acompaña a cada caso corresponde a su número de segmentación generado. Asimismo, se cuantifica el tiempo de inferencia para segmentar la hoja y su consumo de memoria *RAM*; las cifras menores se resaltan en color azul.

Tabla 7.54. Evaluación cualitativa y cuantitativa del funcionamiento.

a) (1)		b) (1)		c) (1)	
					
Tiempo (ms)	Memoria (Mbytes)	Tiempo (ms)	Memoria (Mbytes)	Tiempo (ms)	Memoria (Mbytes)
2,762	45	2,830	68	2,546	35
a) (2)		b) (2)		c) (2)	
					
Tiempo (ms)	Memoria (Mbytes)	Tiempo (ms)	Memoria (Mbytes)	Tiempo (ms)	Memoria (Mbytes)
2,695	110	2,839	98	2,634	93

El caso “c) (1)” registró los indicadores más bajos en tiempo de inferencia y uso de la memoria *RAM* del sistema móvil, logrando una segmentación casi perfecta puesto que, se observa algunas partículas minúsculas no pertenecientes al área foliar; los casos “a) (1)” y “b) (1)” presentan las mismas características.

Retomando el caso b) de la Tabla 7.38 donde se observa una impresión de localización a simple vista, *MobileSAM* en conjunto con *YoloLeaf* demostraron un correcto desempeño al separar las áreas foliares acertadamente ilustrado en los casos “b) (1)” y “b) (2)” (Tabla 7.54), a través de los cuadros delimitadores recibidos como *prompt* de entrada.

CAPÍTULO 8

CONCLUSIONES

El capítulo expone las conclusiones principales del trabajo, la verificación del cumplimiento de los objetivos y los alcances logrados, así como las perspectivas de trabajo futuro derivadas de la investigación. Adicionalmente, se anexan las actividades académicas complementarias realizadas durante el programa doctoral.

8.1 Conclusiones generales

La localización de objetivos en entornos agrícolas no estructurados presenta una elevada complejidad debido a la alta variabilidad biológica, en este caso de las hojas de la planta de jitomate, ya que al ser parte de un ente vivo están sujetas a modificaciones fisiológicas inducidas por fitopatologías, deficiencias nutricionales y estrés abiótico, etc., lo que genera considerables variaciones morfológicas, dimensionales y cromáticas entre distintas especies y estadios de desarrollo.

Adicionalmente, las plagas de insectos alteran las características visuales de una hoja mediante lesiones, rupturas, presencia de estructuras ajenas a la planta como telarañas, capullos, etc. Estas alteraciones provocan oclusiones parciales o totales, anomalías en la coloración y deformaciones severas del contorno, aumentando la complejidad del problema para localizar el objetivo y con ello se incrementa la tasa de falsos negativos en la tarea de detección.

Para mitigar estas limitaciones, la metodología propuesta integra una arquitectura dual basada en dos modelos de aprendizaje profundo especializados en tareas complementarias: detección de objetos y segmentación de instancias. Este enfoque combinado reduce significativamente las imprecisiones en la delimitación foliar sobre imágenes adquiridas bajo condiciones de campo abierto.

El modelo propuesto *YoloLeaf*, redujo el número de parámetros de entrenamiento en 26.02%, 26.75%, y el espacio de almacenamiento en disco en 25.71%, 25.00% comparado con los modelos *Yolov12n* y *Yolov11n* respectivamente. En términos de operaciones de punto flotante (*FLOPS*) también se registró una disminución del 7.93% en ambos modelos comparativos.

Al evaluar el modelo *YoloLeaf* con el conjunto de validación, *YoloLeaf* redujo la tasa de falsos negativos en un 4.34 % respecto a *Yolov12n* y un 2.50 % respecto a *Yolov11n*, lo que se traduce en un incremento en la exhaustividad (*recall*) y una menor omisión de láminas foliares en los 10 escenarios evaluados. Además, se minimizaron los errores del trazado de los cuadros delimitadoras en 0.11 y 0.30 puntos porcentuales de *mAP@50* y *mAP@50-95* en comparación con *Yolov11n*, y en 1.02 y 1.08 puntos frente a *Yolov12n*.

Adicionalmente, se cuantificó la capacidad de generalización de los modelos *Yolov12n*, *Yolov11n* y *YoloLeaf* utilizando conjuntos de datos de otros cultivos, como cacahuete, frijol

común y frijol rojo. El modelo propuesto mostró una capacidad superior de inferencia al alcanzar las métricas más altas: 92.00 % de precisión (*Precision*), 88.70 % de exhaustividad (*Recall*), 93.70 % de *mAP@50* y 86.70 % de *mAP@50-95*. En contraste, sobre las imágenes de frijol rojo capturadas en perspectiva de 90° durante la cuarta semana de desarrollo, *Yolov11n* y *Yolov12n* alcanzaron una precisión del 86.00 % y 87.00 %, un *Recall* del 83.70 % y 83.10 %, un *mAP@50* del 91.40 % en ambos casos, y valores de *mAP@50-95* del 75.40 % y 73.50 %, respectivamente.

En ambos casos de estudio, *YoloLeaf* minimizó las incidencias de localización de hojas con respecto a *Yolov12n* y *Yolov11n*, proporcionando un mayor número de regiones delimitadas (*Bounding Boxes*) válidas, las cuales se emplean posteriormente como *prompts* de entrada para la segmentación con los modelos *SAM2-Tiny* y *MobileSAM*.

a) Objetivo general

Tabla 8.1. Cumplimiento del objetivo general.

Objetivo	Descripción
Modificar la arquitectura de una red neuronal para localizar y segmentar las hojas de las plantas de jitomate en imágenes capturadas a campo abierto.	Se creó una nueva variante (<i>YoloLeaf</i>) basada en <i>Yolov11n</i> que reduce la probabilidad de falsos negativos al calcular la ubicación y extensión espacial de láminas foliares existentes en una imagen. A partir de los cuadros delimitadores generados por <i>YoloLeaf</i> , los modelos <i>SAM2-Tiny</i> y <i>MobileSAM</i> son empleados para delimitar el área foliar, aprovechando el potencial del paradigma <i>zero-shot learning</i> .

b) Objetivos específicos

Tabla 8.2. Cumplimiento de los objetivos específicos.

Objetivos	Descripción
Analizar el estado del arte relacionado con el tema de detección de hojas y frutos en entornos agrícolas naturales, mediante aprendizaje profundo, con el propósito de identificar las metodologías, el desempeño y limitaciones de las actuales arquitecturas de redes.	Se llevó un análisis de artículos relacionados con el tema de la agricultura de precisión basada en visión por computadora, centrados en tareas de detección y segmentación sobre imágenes agrícolas. Se examinaron las siguientes secciones: I) las propuestas de solución, II) materiales y recursos utilizados, III) métricas de evaluación, IV) producto final (contribuciones) y V) trabajos a futuro.
Construir un conjunto de imágenes de plantas de jitomate etiquetadas mediante un protocolo de anotación para generar el <i>ground truth</i> requerido en las etapas de entrenamiento y validación de los modelos de redes (original y propuesto).	Se recolectaron imágenes de plantas de jitomate capturadas bajo diversas condiciones complejas en exteriores, a través de visitas a diferentes sitios web como: blogs, foros, redes sociales y repositorios. Posteriormente, se realizó la construcción del <i>ground truth</i> mediante el etiquetado de manera manual de las instancias delimitadoras.
Desarrollar modificaciones en la arquitectura seleccionada para mejorar la capacidad de	Se realizaron dos modificaciones en la arquitectura: I) la sustitución de módulos base por un bloque compuesto por operaciones convolucionales ligeras y dos mecanismos

Objetivos	Descripción
generalización del modelo original en entornos complejos del cultivo de las plantas de jitomate.	de atención basados en Vision Transformers (ViT) para afinar los mapas de características,y, II) el reemplazo de la función de pérdida <i>CloU</i> por <i>WIoUv3</i> para mejorar la localización de los objetos en la imagen.
Validar de forma cuantitativa y cualitativa el desempeño de la red propuesta, empleando métricas clásicas de evaluación para contrastar los resultados frente al modelo base y a soluciones competitivas del estado del arte.	Se realizó un análisis comparativo entre la arquitectura propuesta <i>YoloLeaf</i> , el modelo de referencia <i>Yolov11n</i> y la versión más reciente, <i>Yolov12n</i> (publicada en 2025). La evaluación se estructuró mediante métricas estandarizadas en dos dimensiones principales: I) la eficiencia del modelo, determinada por el tamaño de almacenamiento (MB) y la complejidad computacional (FLOPs); y II) el rendimiento predictivo, evaluado a través de <i>Precision</i> , <i>Recall</i> , <i>F1-Score</i> , <i>mAP@50</i> y <i>mAP@50-95</i> .»

c) Alcances

Tabla 8.3. Cumplimiento de los alcances.

Alcance	Descripción
Las imágenes exhiben hojas de la planta del jitomate saludables y anómalas capturados en ambientes naturales.	Se recolectaron diversas imágenes digitales que muestran plantas de jitomate en diferentes condiciones. Las muestras presentan diferentes tipos de forma (tipo papa y sierra) en condiciones aparentemente saludables y anómalas (enfermedad, deficiencias de nutrientes, estrés, destrucción etc.). Las tomas se realizaron en entornos naturales no controlados, considerando variaciones en la iluminación, múltiples perspectivas, diversos ángulos de captura y presencia de oclusiones.
Se realiza el análisis de imágenes de las hojas del jitomate en sus etapas de desarrollo y maduración.	Se incorporaron imágenes que muestran el cultivo de jitomate en sus fases vegetativa, de floración y de fructificación.
Se localiza y segmenta la mayor cantidad de hojas de diferentes especies del cultivo de interés.	El modelo propuesto, <i>YoloLeaf</i> , registró un desempeño superior en la localización de hojas que presentan variaciones morfológicas de forma y tamaño, al reducir el número de falsos negativos en comparación con <i>Yolov11n</i> y <i>Yolov12n</i> .
Se segmenta tanto el haz (parte frontal) como el envés (parte trasera) de la lámina foliar.	Durante la experimentación el modelo logró segmentar: I) áreas parciales (oclusiones entre hojas y otros objetos), II) formas irregulares de la lámina foliar y III) superficies dañadas (agujeros).
Se trabaja con ejemplares digitales de conjunto públicos disponibles y recolección a través de visitas a sitios web en internet del cultivo de interés, fotografiados en condiciones adversas.	Se hizo una recolección de 1,562 imágenes a través de diferentes sitios web y 1,438 provenientes del repositorio <i>RF</i> , tomadas bajo entornos naturales no controlados (cambios de iluminación, diferentes perspectivas y ángulos de captura, presencia de oclusiones, etc.).
Se utiliza herramientas en línea para elaborar el etiquetado de instancias y el entrenamiento del modelo.	Se recurrió a los insumos gratuitos disponibles a través del internet: I) <i>RF</i> repositorio público de imágenes, II) <i>CVAT</i> aplicación web para elaborar el etiquetado de instancias, III) <i>GoogleColab</i> servidor virtual para el entrenamiento, generación y evaluación de <i>YoloLeaf</i> .

Alcance	Descripción
	Cabe mencionar que <i>GoogleColab</i> es gratuito las primeras 3 horas de uso. El tiempo se restablece cada 24 horas, por lo que se pagó el plan " <i>GoogleColabPro</i> " para disponer de más tiempo para realizar múltiples entrenamientos.

d) Contribución científica

- Se entrega un modelo enfocado en tareas de localización de hojas de plantas en una versión compacta (menos espacio de almacenamiento) y mejorada (menor probabilidad de casos falsos negativos). El modelo se integra de modificaciones estructurales a la arquitectura de red *Yolov11n* en dos aspectos: optimizar la extracción de características y reducir la carga computacional en *FLOPs*. Para mejorar la extracción de características el componente *MobileViTBlock* reemplaza a *C3k2* y *C2PSA* mediante la fusión de características locales y globales calculadas por las operaciones convolucionales y por los transformadores de visión respectivamente. Para disminuir las operaciones de punto flotante se removieron los módulos originales: *Conv*, *C3k2* y *C2PSA* ubicados en los índices siete, ocho y diez del *backbone*, respectivamente y se reemplazaron por los módulos de *MV2Block* y *MobileViTBlock*.

e) Contribución de datos y recursos

- Se proporciona un conjunto limpio de 3,000 imágenes para ser utilizado en tareas de detección de hojas de plantas de jitomate. Las muestras representan la diversidad morfológica foliar del cultivo (morfotipos de lámina tipo papa y dentada/aserrada), registrando ejemplares con estado aparentemente saludables y con anomalías fisiológicas asociadas a estrés biótico (enfermedades y plagas) y abiótico (deficiencias nutrimentales y daño por estrés, destrucción, etc.). Las tomas, presentan variaciones en la iluminación, múltiples perspectivas, diversos ángulos de captura y presencia de oclusiones.
- Se construyó el *ground truth* del conjunto de imágenes integrado por el etiquetado de 27,397 instancias usando los cuadros delimitadores tradicionales.

f) Contribución tecnológica

- Se desarrolló una aplicación móvil que implementa *YoloLeaf* para localizar el objetivo y *MobileSAM* para segmentar el objeto. Se otorga el código fuente del sistema para futuras mejoras.

8.2 Actividades Académicas

- Participación en la presentación de posters en la escuela de inteligencia computacional y robótica, celebrado del 16 al 20 de agosto de 2022, Emiliano Zapata, Morelos, México.



- II. Participación como ponente en la conferencia “Proyectos de investigación de la línea de investigación de inteligencia artificial”, celebrado el día 29 de marzo de 2023, Cuautla, Morelos, México.



- III. Participación en el programa ENC 2023 (Encuentro Nacional de Computación) celebrado del 11 al 15 de septiembre de 2023, Guanajuato, México.



- IV. Asistencia en el taller nacional de drones inteligentes celebrado el 5 y 6 de octubre de 2023, San Andrés, Cholula, Puebla, México.



- V. Concurso de exposición de poster en la Escuela de Inteligencia Computacional y Robótica 2023 celebrado del 10 al 14 de octubre de 2023, Emiliano Zapata, Morelos, México.



- VI. Presentación del artículo titulado “Segmentación de imágenes de entornos naturales para la agricultura de precisión” en el evento CITCA 2024, celebrado del 19 al 22 de noviembre de 2024, Cuernavaca, Morelos, México.



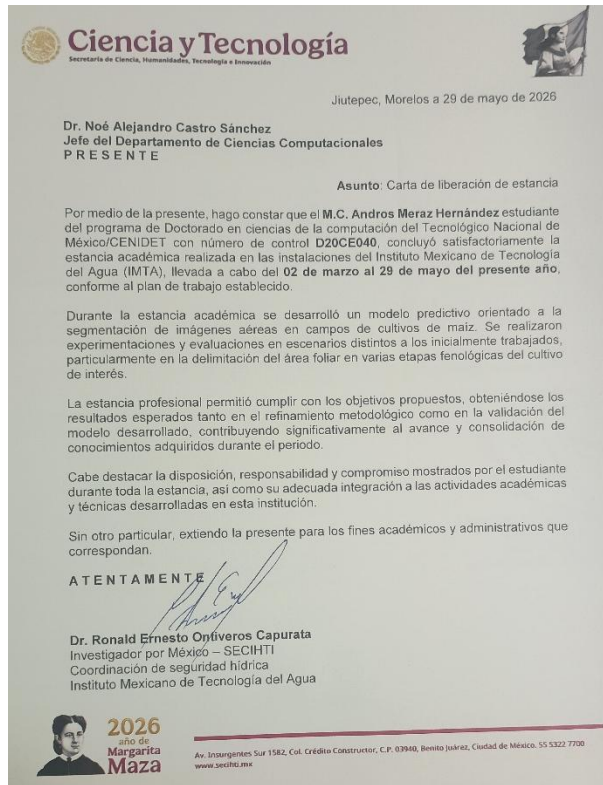
- VII. Publicación del artículo (contribución de esta investigación) “*Deep Learning Architecture for Tomato Plant Leaf Detection in Images Captured in Complex Outdoor Environments*” en la revista *Mathematics* el 22 de julio de 2025.



- VIII. Participación como jurado en el “Concurso Nacional de Prototipos y Proyectos de emprendimiento 2026”, en el CETIs 99, ubicado en Yautepec, Morelos, México, celebrado el 28 de noviembre de 2025.



IX. Estancia profesional en el Instituto Mexicano de Tecnología del Agua (IMTA), Jiutepec, Morelos, México, del 02 de marzo al 29 de mayo de 2026.



8.3 Trabajo a futuro

Debido al elevado número de falsos positivos registrados por el modelo *YoloLeaf* se propone modificar la entrada original de tres canales *RGB* mediante la integración de canales adicionales de información, tales como mapas de profundidad, índices de vegetación e imágenes *NIR* (*Near-Infrared*) reales o sintéticas.

InfraGAN (Özkanoğlu & Özer, 2022) es una arquitectura de Redes Neuronales Adversarias (GAN) diseñada específicamente para la traducción de dominios de imágenes, permitiendo convertir imágenes del espectro visible (*RGB* o escala de grises) al espectro infrarrojo (*IR*). Su objetivo principal es la creación automática de datos sintéticos de alta fidelidad en situaciones cuando no existen conjuntos de datos reales.

Por ejemplo, en (Liu *et al.*, 2023) se enfocan en identificar con precisión los tallos principales de las plantas de tomate. Se propone un modelo llamado *YOLACTFusion* que hace la combinación de imágenes a color (*RGB*) con imágenes de infrarrojo cercano (*NIR*) de entre 900 y 1100 *nm* para reducir los fallos que presentan los robots al identificar el objetivo de interés, ya que estos tienen un color muy similar al de las hojas y el fondo complejo. Los autores reportan mejora en la métrica *mAP* del 39.20% al 46.29% comparado con el modelo base *YOLACT*.

Depth Anything es un modelo diseñado para inferir la estimación de profundidad monocular del objetivo a partir del análisis de sus principales características como volumen, forma, texturas etc. logrando separar con precisión dónde termina un objeto y dónde empieza el fondo, incluso si tienen colores muy parecidos (Yang *et al.*, 2024); *Depth Anything V3* publicado en 2025 es la versión más reciente.

En (Nan *et al.*, 2023), se propone una estrategia para aislar las hojas de la planta del maíz que figuran en imágenes capturadas en entornos naturales complejos. Se emplean cámaras *RGB-D* para capturar la estructura de la planta. Utilizando la información de los mapas de profundidad, se desarrolla un algoritmo de umbralización espacial que elimina aquellos píxeles que se encuentren fuera del rango de distancia del plano principal de la planta, lo que permite una extracción del objeto a partir de los bordes de la lámina foliar.

En conjunto podrían reducir el número de falsos positivos observado durante las pruebas al separar las tonalidades de color verde (orgánicas y no orgánicas) y determinar una distancia que favorezca a un correcto desempeño del modelo.

Referencias

- Abdullah, A.; Amran, G.A.; Tahmid, S.M.A.; Alabrah, A.; AL-Bakhrani, A.A.; Ali, A. (2024). A Deep-Learning-Based Model for the Detection of Diseased Tomato Leaves. DOI: <https://doi.org/10.3390/agronomy14071593>
- Abulizi A, Ye J, Abudukelimu H and Guo W. (2025). DM-YOLO: improved YOLOv9 model for tomato leaf disease detection. *Front. Plant Sci.* 15:1473928. doi: 10.3389/fpls.2024.1473928
- Alexopoulos, C. J., Mims, C. W., & Blackwell, M. (1996). *Introductory Mycology* (4^a ed.). John Wiley & Sons.
- Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby. (2021). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Published as a conference paper at ICLR 2021. <https://doi.org/10.48550/arXiv.2010.11929>
- Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, Choong Seon Hong. (2023). Faster Segment Anything: Towards Lightweight SAM for Mobile Applications. <https://doi.org/10.48550/arXiv.2306.14289>
- Computer Vision Annotation Tool (CVAT) (s.f). Recuperado 6 de junio 2025, de <https://app.cvat.ai/auth/login?next=/jobs>
- Chollet, F. (2018). *Deep learning with Python*. Shelter Island, NY: Manning Publications.
- Chollet, F. (2021). *Deep Learning with Python* (2^a ed.). Manning Publications.
- D.C. L. Márquez. (2021). “Sobresegmentación de Imágenes a Color Mediante un Grafo Difuso Basado en una Medida de Distancia Espacial y Color”. Tesis CENIDET.
- D.G.S. Santiago. (2018) “Sintonización de una Red Totalmente Conectada para Segmentación de Dos Clases de Objetos en Imágenes”. Tesis CENIDET.
- Delin Wu, Shan Jiang, Enlong Zhao, Yilin Liu, Hongchun Zhu, Weiwei Wang, Rongyan Wang. 2022. Detection of Camellia oleifera Fruit in Complex Scenes by Using YOLOv7 and Data Augmentation. DOI: <https://doi.org/10.3390/app122211318>
- Esau, K. (1977). Las hojas. En *Anatomía de las Plantas con Semilla* (3^a ed., pp. 200–250). Omega.
- Gad, A. F. (2020, 16 de octubre). Evaluating object detection models using mean Average Precision (mAP). Paperspace Blog. <https://blog.paperspace.com/mean-average-precision/>
- Gary Storey, Qinggang Meng, Baihua Li. 2022. Leaf Disease Segmentation and Detection in Apple Orchards for Precise Smart Spraying in Sustainable Agriculture. DOI: <https://doi.org/10.3390/su14031458>
- Grandini, M., Bagli, E. and Visani, G. (2020) 'Metrics for Multi-Class Classification: an Overview', arXiv. DOI: <https://doi.org/10.48550/arXiv.2008.05756>.
- [Google Colab](https://colab.google/). (s.f). Recuperado 8 de abril de 2024, de <https://colab.google/>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

Hamish A. Craze, Nelishia Pillay, Fourie Joubert, Dave K. Berger. 2022. Deep Learning Diagnostics of Gray Leaf Spot in Maize under Mixed Disease Field Conditions. DOI: <https://doi.org/10.3390/plants11151942>

Hicham Slimani, Jamal El Mhamdi, Abdelilah Jilbab. (2023). Artificial Intelligence-based Detection of Fava Bean Rust Disease in Agricultural Settings: An Innovative Approach. International Journal of Advanced Computer Science and Applications(IJACSA), 14(6). DOI: <http://dx.doi.org/10.14569/IJACSA.2023.0140614>

Hughes, D. P., & Salathé, M. (2015). *An open access repository of images on plant health to enable the development of mobile disease diagnostics through machine learning and crowdsourcing*. arXiv. <https://arxiv.org/abs/1511.08060>

Ilyas, T., Arsa, D. M. S., Ahmad, K., Lee, J., Won, O., Lee, H., Kim, H., & Park, D. S. (2025). CWD30: A new benchmark dataset for crop-weed recognition in precision agriculture. *Computers and Electronics in Agriculture*, 229, 109737. <https://doi.org/10.1016/j.compag.2024.109737>

ImageNet. (s.f.). Recuperado 8 de febrero 2022, de <http://image.net.org/challenges/LSVRC/2012/>

INIFAP. (2024). *El cacahuete: una opción de cultivo en tierras de temporal*. Gobierno de México. <https://www.gob.mx/inifap/articulos/el-cacahuete-una-opcion-de-cultivo-en-tierras-de-temporal>

Hu, W.; Hong, W.; Wang, H.; Liu, M.; Liu, S. (2023). "A Study on Tomato Disease and Pest Detection Method". *Appl. Sci.* 2023, 13, 10063. <https://doi.org/10.3390/app131810063>

Jia Yao, Yubo Wang, Ying Xiang, Jia Yang, Yuhang Zhu, Xin Li, Shuangshuang Li, Jie Zhang, Guoshu Gong. 2022. Two-Stage Detection Algorithm for Kiwifruit Leaf Diseases Based on Deep Learning. DOI: <https://doi.org/10.3390/plants11060768>

Kaur, R., Mittal, U., Wadhawan, A. *et al.* YOLO-LeafNet: a robust deep learning framework for multispecies plant disease detection with data augmentation. *Sci Rep* **15**, 28513 (2025). <https://doi.org/10.1038/s41598-025-14021-z>

Keiron O'Shea & Ryan Nash. (2015). An Introduction to Convolutional Neural Networks. DOI: <https://doi.org/10.48550/arXiv.1511.08458>

Khalid, S.; Oqaibi, H.M.; Aqib, M.; Hafeez, Y. (2023). Small Pests Detection in Field Crops Using Deep Learning Object Detection. *Sustainability* 15, 6815. <https://doi.org/10.3390/su15086815>

Khan F, Zafar N, Tahir MN, Aqib M, Waheed H and Haroon Z (2023) A mobile based system for maize plant leaf disease detection and classification using deep learning. *Front. Plant Sci.* 14:1079366. DOI: 10.3389/fpls.2023.1079366

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., Dollár, P., y Girshick, R. (2023). Segment Anything. arXiv. doi.org

Li Xudong, Zhou Yuhong, Liu Jingyan, Wang Linbai, Zhang Jun, Fan Xiaofei. 2022. The Detection Method of Potato Foliage Diseases in Complex Background Based on Instance Segmentation and Semantic Segmentation. *Frontiers in Plant Science*, vol 13. DOI: 10.3389/fpls.2022.899754

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018). *Focal loss for dense object detection*. arXiv. <https://arxiv.org/abs/1708.02002>

Liu, C., Feng, Q., Sun, Y., Li, Y., Ru, M., & Xu, L. (2023). YOLACTFusion: An instance segmentation method for RGB-NIR multimodal image fusion based on an attention mechanism. *Computers and Electronics in Agriculture*, 213, Artículo 108186. <https://doi.org/10.1016/j.compag.2023.108186>

Liu, W.; Zhai, Y.; Xia, Y. Tomato Leaf Disease Identification Method Based on Improved YOLOX. *Agronomy* 2023, 13, 1455. <https://doi.org/10.3390/agronomy13061455>

Liu,Z.; Guo,X.;Zhao,T.; Liang, S. YOLO-BSMamba: A YOLOv8s-BasedModelforTomato Leaf Disease Detection inComplex Backgrounds. *Agronomy*2025,15,870. <https://doi.org/10.3390/agronomy15040870> 27,557bb

López Aguilar, R. L., Rodríguez Ramírez, M. R., & Valdez Flores, E. V. (2004). *Cultivo de tomate en el valle Vizcaíno*. Centro de Investigaciones Biológicas del Noroeste, S.C. Recuperado de: <https://cibnor.repositorioinstitucional.mx/jspui/bitstream/1001/1319/1/LIBRO%20CULTIVO%20DE%20TOMATE%20EN%20EL%20VALLE%20DEL%20VIZCAINO.pdf>

Lucas, J. A. (2020). *Plant Pathology and Plant Pathogens* (4^a ed.). Wiley-Blackwell.

Lucas M. Tassis, João E. Tozzi de Souza, Renato A. Krohling. 2021. A deep learning approach combining instance and semantic segmentation to identify diseases and pests of coffee leaves from in-field images. DOI: <https://doi.org/10.1016/j.compag.2021.106191>

Matthews, R. E. F. (2012). *Fundamentals of Plant Virology*. Academic Press.

Mario Parreño L. (2018) Exploración de topologías para el reconocimiento de género [Tesis de Maestría, Universidad Politécnica de Valencia]. <http://hdl.handle.net/10251/115338>

MathWorks. (s.f.). Ground Truth. Recuperado el 8 de noviembre de 2022, de <https://la.mathworks.com/discovery/ground-truth.html>

Meraz H.A. (2022). “Sistema de visión artificial para la detección de plantas enfermas mediante aprendizaje profundo”. Tesis CENIDET.

Meraz-Hernández, A., Fuentes-Pacheco, J., Magadán-Salazar, A., Pinto-Elías, R., & González-Franco, N. (2025). Deep Learning Architecture for Tomato Plant Leaf Detection in Images Captured in Complex Outdoor Environments. *Mathematics*, 13(15), 2338. <https://doi.org/10.3390/math13152338>

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. Chen. (2018). "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 2018 pp. 4510-4520. doi: 10.1109/CVPR.2018.00474

Mingfeng Huang, Guoqin Xu, Junyu Li and Jianping Huang. (2021). A Method for Segmenting Disease Lesions of Maize Leaves in Real Time Using Attention YOLACT++. DOI: <https://doi.org/10.3390/agriculture11121216>

Moupojou, E., Tagne, A., Retraint, F., Tadonkemwa, A., Dongmo, W., Tapamo, H., & Nkenlifack, M. (2023). FieldPlant: A dataset of field plant images for plant disease detection and classification with deep learning. *IEEE Access*, 11, 35398–35410. DOI: <https://doi.org/10.1109/ACCESS.2023.3263042>

M.M. Xicohtencatl. (2018). “Segmentación no paramétrica de Tejidos Cerebrales Mediante una Arquitectura Paralela de Redes Neuronales Convolucionales”. Tesis CENIDET.

MSU (s.f.) Información de enfermedades del jitomate. Available at: [http://extension.msstate.edu/publications/common-diseases tomatoes](http://extension.msstate.edu/publications/common-diseases-tomatoes) (Accessed: 26 September 2025).

Nan, F., Song, Y., Yu, X., Nie, C., Liu, Y., Bai, Y., Zou, D., Wang, C., Yin, D., Yang, W., & Jin, X. (2023). A novel method for maize leaf disease classification using the RGB-D post-segmentation

image data. *Frontiers in Plant Science*, 14, Artículo 1268015. <https://doi.org/10.3389/fpls.2023.1268015>

Özkanoglu, M. A., & Özer, S. (2022). InfraGAN: A structural similarity based GAN for infra-red image synthesis. *IEEE Access*, 10, 33113-33123. DOI: <https://doi.org/10.1016/j.patrec.2022.01.026>

Parez, S.; Dilshad, N.; Alghamdi, N.S.; Alanazi, T.M.; Lee, J.W. Visual Intelligence in Precision Agriculture: Exploring Plant Disease Detection via Efficient Vision Transformers. *Sensors* 2023, 23, 6949. <https://doi.org/10.3390/s23156949>

Perry, R. N., & Moens, M. (Eds.). (2013). *Plant Nematology* (2^a ed.). CABI

Rashid, J., Khan, I., Ali, G., Rehman, S. U., Alturise, F., & Alkhalifah, T. (2023). Real-time multiple guava leaf disease detection from a single leaf using hybrid deep learning technique. *Computers, Materials & Continua*, 74(1), 1235–1257. <https://doi.org/10.32604/cmc.2023.032005>

Rahima Khanam and Muhammad Hussain. (2024). “YOLOV11: AN OVERVIEW OF THE KEY ARCHITECTURAL ENHANCEMENTS”. DOI: <https://doi.org/10.48550/arXiv.2410.17725>

Rahmadian, F., Yunandra, Y., Zuhry, E., Irfandri, I., Saputra, R., Masruri, N. W., Br. Hutabarat, Y. A., Simanjuntak, N. Y., & Destiani, E. (2024). Characterization And Genetic Relationship of Several Tomato (*Solanum lycopersicum* L.) Genotypes from Various Regions. *JURNAL AGRONOMI TANAMAN TROPIKA (JUATIKA)*, 6(2), 208–223. <https://doi.org/10.36378/juatika.v6i2.3579>

Ravi, N., Gabeur, V., Hu, Y.-T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K. V., Carion, N., Wu, C.-Y., Girshick, R., Dollár, P., y Feichtenhofer, C. (2024). SAM 2: Segment Anything in Images and Videos. arXiv. doi.org

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You Only Look Once: Unified, Real-Time Object Detection. arXiv. <https://doi.org/10.48550/arXiv.1506.02640>

Risa Shinoda, Hirokatsu Kataoka, Kensho Hara, Ryozi Noguchi. 2023. Transformer-based ripeness segmentation for tomatoes. *Smart Agricultural Technology*, Volume 4. DOI: <https://doi.org/10.1016/j.atech.2023.100196>

Roboflow. (s.f). Recuperado 8 de abril de 2024, de <https://universe.roboflow.com/>

Russell, S., & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4^a ed.). Pearson.

Sachin Mehta, Mohammad Rastegari. (2022). MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer. Published as a conference paper at ICLR 2022. <https://doi.org/10.48550/arXiv.2110.02178>

Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., y Chen, L.-C. (2019). MobileNetV2: Inverted Residuals and Linear Bottlenecks. DOI: <https://arxiv.org/abs/1801.04381>

Sanghyun Woo, Jongchan Park, Joon-Young Lee, In So Kweon. (2018). CBAM: Convolutional Block Attention Module. DOI: <https://doi.org/10.48550/arXiv.1807.06521>

Secretaría de Agricultura y Desarrollo Rural (SAGARPA) (2017). Planeación agrícola nacional 2017-2030. Jitomate mexicano. Recuperado de <https://www.gob.mx/cms/uploads/attachment/file/257077/Potencial-Jitomate.pdf>

Secretaría de Agricultura y Desarrollo Rural (SAGARPA) (2017). *FRIJOL Mexicano: Monografía del cultivo* [PDF]. https://www.gob.mx/cms/uploads/attachment/file/256428/B_sico-Frijol.pdf

Secretaría de Agricultura y Desarrollo Rural. (2023, October 13). *El cacahuete en México: tesoro nutricional de la agricultura mexicana*. Gobierno de México. <https://www.gob.mx/agricultura/articulos/el-cacahuete-en-mexico-tesoro-nutricional-de-la-agricultura-mexicana>

Singh, D., Jain, N., Jain, P., Kayal, P., Kumawat, S., & Batra, N. (2020). *PlantDoc: A dataset for visual plant disease detection*. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD* (pp. 249–253). ACM. DOI: <https://doi.org/10.1145/3371158.3371196>

Sonka, M. and Fitzpatrick, J.M. (2009) *Handbook of Medical Imaging. Volume 2: Medical Image Processing and Analysis*. Bellingham, WA: SPIE.

Tao, J.; Li, X.; He, Y.; Islam, M.A. CEFW-YOLO: A High-Precision Model for Plant Leaf Disease Detection in Natural Environments. *Agriculture* 2025, 15, 833. <https://doi.org/10.3390/agriculture15080833>

Schumann, G. L., & D'Arcy, C. J. (2012). *Essential Plant Pathology* (2^a ed.). American Phytopathological Society (APS Press).

Szeliski, R. (2022). *Computer Vision: Algorithms and Applications* (2^a ed.). Springer.

Tao, J., Li, X., He, Y., & Islam, M. A. (2025). CEFW-YOLO: A high-precision model for plant leaf disease detection in natural environments. *Agriculture*, 15(8), 833. <https://doi.org/10.3390/agriculture15080833>

Tian, Y., Ye, Q., & Doermann, D. (2025). *YOLOv12: Attention-centric real-time object detectors*. arXiv. <https://arxiv.org/abs/2502.12524>

Tong, Z., Chen, Y., Xu, Z., & Yu, R. (2023). *Wise-IoU: Bounding box regression loss with dynamic focusing mechanism*. arXiv. <https://arxiv.org/abs/2301.10051>

Ultralytics. (s.f). Recuperado 8 de abril de 2024, de <https://docs.ultralytics.com/guides>

Üzen, H., Türkoğlu, M., Aslan, M., & Hanbay, D. (2022). Depth-wise Squeeze and Excitation Block-based Efficient-Unet model for surface defect detection. *The Visual Computer*, 39, 1-15. DOI: <https://doi.org/10.3390/s21196565>

Valla, J. J. (2007). *Botánica: Morfología de las Plantas Superiores* (2^a ed.). Hemisferio Sur.

Wang X and Liu J (2025) TomatoGuard-YOLO: a novel efficient tomato disease detection method. *Front. Plant Sci.* 15:1499278. doi: 10.3389/fpls.2024.1499278

Wang, G., Chen, Y., An, P., Hong, H., Hu, J., Huang, T. (2023). "UAV-YOLOv8: A Small-Object-Detection Model Based on Improved YOLOv8 for UAV Aerial Photography Scenarios". *Sensors* 2023, 23, 7190. <https://doi.org/10.3390/s23167190>

Wang Y, Zhang P and Tian S (2024) Tomato leaf disease detection based on attention mechanism and multi-scale feature fusion. *Front. Plant Sci.* 15:1382802. DOI: 10.3389/fpls.2024.1382802

Xiao, H., Ning, C., He, H., et al. (2024). *YOLOv8-GK: Tomato leaf disease detection algorithm*. Research Square. DOI: <https://doi.org/10.21203/rs.3.rs-5454516/v1>

Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., y Zhao, H. (2024). Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. arXiv. DOI: <https://arxiv.org/abs/2401.10891>

Yassine Lakhdari, Enric Soldevila, Jihene Rezgui. (2023). Detection of plant diseases in an industrial greenhouse: Development, Validation & Exploitation. *Laboratoire Recherche Informatique Maisonneuve (LRIMa)*, Montreal, Canada. IEEE_ISNCC_2023-2.pdf (cmaisonneuve.qc.ca)

Yi Xia, Minh Nguyen, Raymond Lutui, Wei Qi Yan. (2023) Multiscale Kiwifruit Detection from Digital Images. Recuperado el 19 noviembre de 2023 de: https://cerv.aut.ac.nz/wp-content/uploads/2023/09/PSIVT_YiXia.pdf

Zhang, C., Han, D., Zheng, S., Choi, J., Kim, T.-H., & Hong, C. S. (2023). *MobileSAMv2: Faster segment anything to everything*. arXiv. <https://arxiv.org/abs/2312.09579>

Zsögön, A., Mateus, V., Reartes, D., & Peres, L. (2017). Understanding and improving water-use efficiency and drought resistance in tomato. En *Achieving sustainable cultivation of tomatoes* (pp. 3-22). Burleigh Dodds Science Publishing. DOI:[10.19103/AS.2016.0007.01](https://doi.org/10.19103/AS.2016.0007.01)