Maestría en Inteligencia Artificial

Visión por Computadora III

Trabajo Final de Investigación

Clasificación Automática de Espacios en Planos Arquitecónicos mediante Segmentación Semántica

Análisis Comparativo de Vision Transformer, U-Net++ y Swin Transformer con Mask R-CNN

Docentes:

- Oksana Bokhonok
- Abraham Adolfo Rodriguez Zepeda

Autores:

- Fabian Sarmiento
- Jorge Cuenca
- Alejandro Lloveras

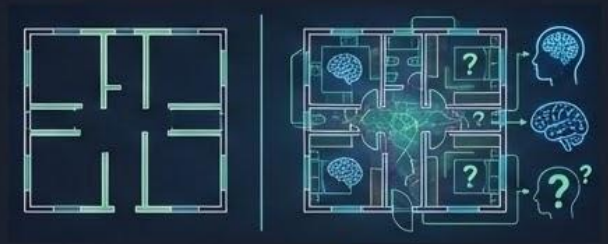
.UBAfiuba
FACULTAD DE INGENIERÍA

Objetivo: Clasificación de Espacios Funcionales

Motivación

- La **clasificación automática** (identificar "cocina" vs. "dormitorio") es crucial para la automatización en arquitectura y análisis urbano.

- **Problema:** Detectar muros es "fácil"; entender la función del espacio requiere comprensión de orden superior.



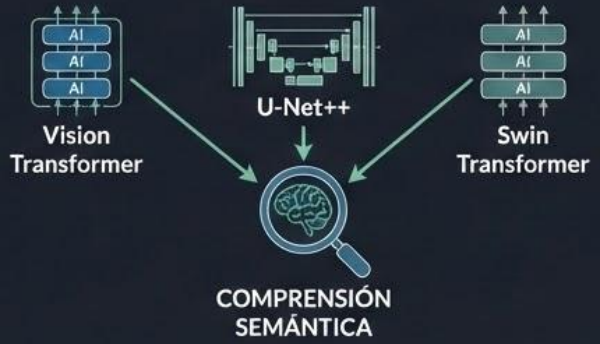
✓ **FÁCIL** (Muros)

✗ **DIFÍCIL** (Comprensión de Orden Superior)

Enfoque Propuesto

- Utilizamos la **segmentación semántica** como tarea intermedia necesaria para extraer límites precisos y características.

- **Objetivo:** Evaluar qué arquitectura moderna facilita mejor esta comprensión semántica.



Dataset CubiCasa5K

Desafíos Técnicos

5000 planos de alta calidad

Variabilidad: diversidad en resolución, estilo de dibujo y densidad.

Gran desbalance de clases:

1. Ratio de desbalance de **456:1.**
2. La clase "Fondo" domina (63.43%), mientras que "Mobiliario Fijo" representa sólo el 0.24%.

Clases reconocidas

Dormitorio	Trastero
Cocina	Cochera
Sala de estar	Lavadero
Baño	Oficina
Comedor	Cuarto de invitados
Pasillo	Cuarto de servicio
Balcón	Otro

Ingeniería de Modelos

Arquitecturas Evaluadas

CNNs vs. Transformers en Tareas Densas

Vision Transformer (ViT): Enfoque base de transformers puros.

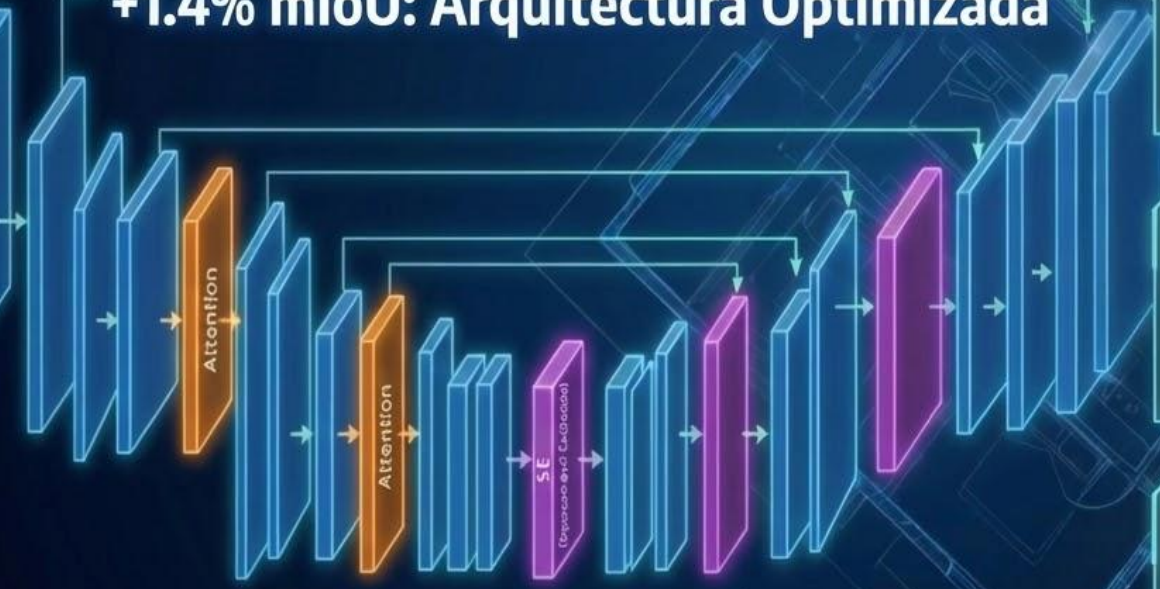
Swin Mask R-CNN: Estado del arte híbrido (*Swin Backbone + Mask R-CNN*).

U-Net++ (Baseline): Evolución de U-Net con conexiones anidadas.

U-Net++ Mejorado (Propio): Nuestra implementación con módulos de atención y recalibración.

U-Net++: Mejoras Arquitectónicas

+1.4% mIoU: Arquitectura Optimizada



U-Net++ Block (Blue) Attention Block (Orange) SE Block (Purple)

Attention Gates
Suprimir regiones irrelevantes (fondo) y enfocarse en estructuras.

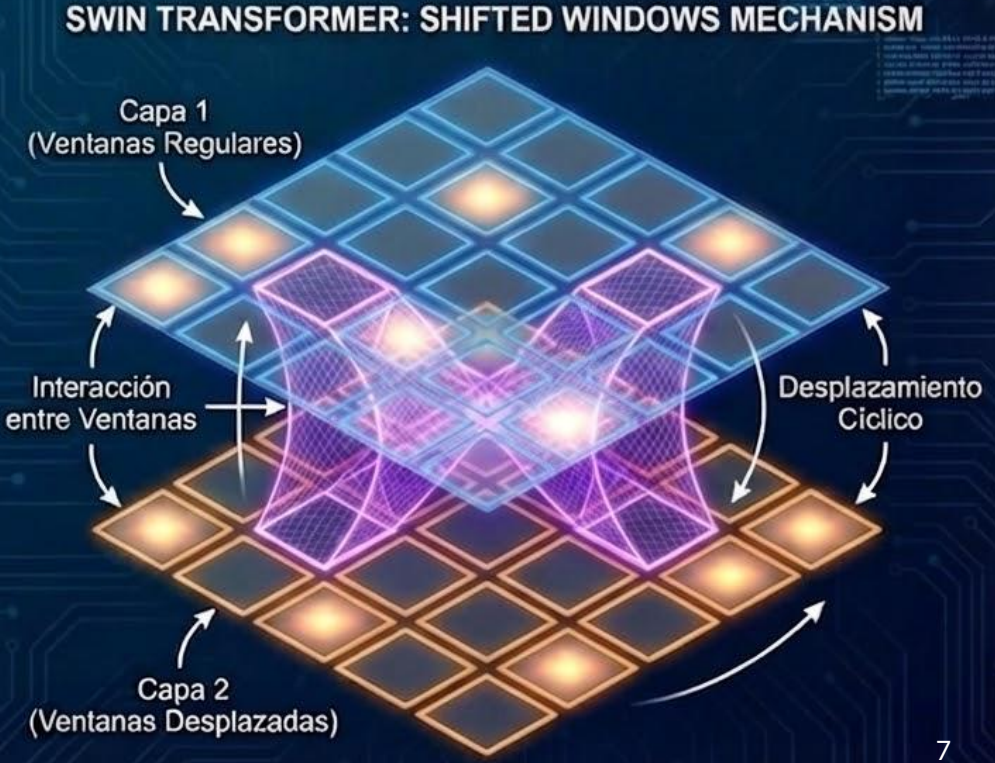
Squeeze-and-Excitation (SE) Blocks
Recalibración adaptativa de los canales de características.

Deep Supervision Mejorada
Facilita la propagación del gradiente y la convergencia.

Swin Transformer + Mask R-CNN:
Mejoras Arquitectónicas

Combinación de modelado global (Transformer) con detección precisa (R-CNN).

- **Backbone Swin:** Usa "Shifted Windows" para calcular atención con complejidad lineal, capturando dependencias a larga distancia.
- **Mask R-CNN Head:** Proporciona segmentación de instancias de alta precisión.



Configuración del Entrenamiento

Función de Pérdida Combinada (Loss Function)

Combatiendo el desbalance de clases

Hardware y Optimizaciones

- **GPU:** 2 x NVIDIA RTX 3090 (24GB VRAM).
- **Técnicas:** Mixed Precision (FP16), Gradient Accumulation Accumulation
- **Aumento de Datos:** Rotación, Color Jitter, Deformación Elástica

Resultados y Métricas

Comparativa de Rendimiento (mIoU)

Los Transformers superan a las CNNs en precisión pura

Mejor rendimiento: *Swin Mask R-CNN*

Mejor tradeoff: *U-Net++ Mejorado*

Las mejoras en U-Net++ lograron un incremento del 3% respecto a la base.

	Escenario de Aplicación	Arquitectura	Justificación
	Tiempo real (>30 FPS)	U-Net++ Base	Alta velocidad, buen rendimiento
	Balance precisión/velocidad	U-Net++ Improved	Mejor trade-off
	Máxima precisión	Swin Mask R-CNN	Rendimiento superior
	Recursos limitados	U-Net++ Base	Menor uso de memoria
	Producción a escala	U-Net++ Improved	Estabilidad y eficiencia

Trade-off: Eficiencia vs. Precisión

Métrica	U-Net++ Mejorado	Swin Mask R-CNN	Conclusión
Parámetros	~11.28 M	87.91 M	Swin es ~8x más pesado
Velocidad	~38 FPS Tiempo Real	12.3 FPS Lento	U-Net++ es tiempo-real
FLOPs	42.3 G	189.5 G	Costo computacional masivo en Swin

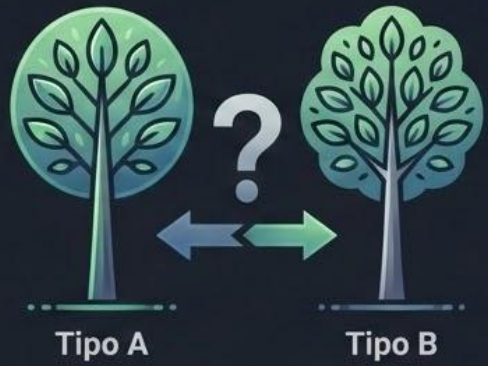
Análisis Cualitativo y Errores

Objetos Pequeños

Swin detecta mejor elementos pequeños gracias a la atención jerárquica.

Límites y Bordes

- U-Net++ tiende a suavizar bordes;
- Swin preserva detalles finos pero puede generar artefactos.



Confusión

Ambos modelos luchan con clases visualmente similares (e.g., tipos de vegetación).

Discusión y Cierre

Aporte de las mejoras implementadas

Efectividad de Mejoras

Los Attention Gates demostraron ser efectivos para focalizar el aprendizaje en estructuras arquitectónicas dispersas.

Superioridad Transformer

La capacidad de modelar dependencias globales (no locales como la convolución) es decisiva en planos complejos.

Robustez

El aumento de datos fue clave para manejar la variabilidad de iluminación y estilo.

Conclusiones Finales

Logros Alcanzados

Swin Mask R-CNN redefine el estado del arte en planos

U-Net++ sigue siendo la reina de la eficiencia.

Se entrega un framework modular y reproducible.

Trabajo Futuro

Explorar arquitecturas híbridas y aprendizaje auto-supervisado para reducir dependencia de etiquetas.

Muchas Gracias

¿Alguna pregunta?

Fabian Sarmiento, Jorge Cuenca & Alejandro Lloveras