

YotoR: You Only Transform One Representation

J. DÍAZ¹,
P. LONCOMILLA²,
J. RUIZ-DEL-SOLAR^{1,2}

¹DIE. Universidad de Chile

²AMTC

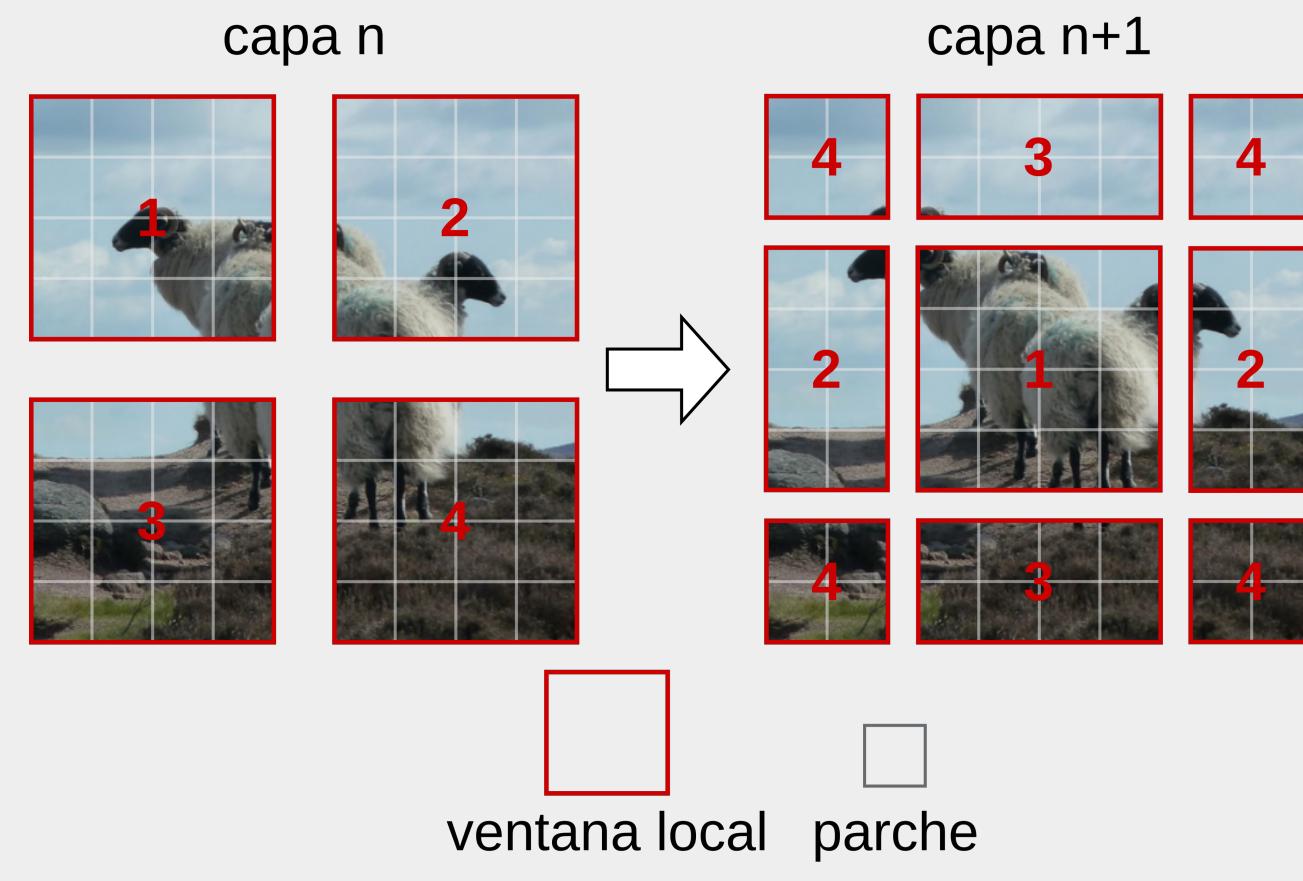


Fig. 1 Partición por ventanas de Swin Transformer

YotoR

- Desarrollamos la arquitectura YotoR al fusionar el **backbone de Swin Transformer** con la **cabeza de YoloR**.
- La idea es aplicar la potente extracción de características del Swin Transformer y usar el conocimiento implícito de YoloR para alinear y mejorar las predicciones.
- El backbone del Swin Transformer está conformado por **cuatro etapas** de bloques Swin Transformer. Esto coincide con el backbone de Darknet de YoloR, lo que simplifica la fusión.
- El **diseño de las conexiones** fue crucial en el desarrollo. Sin una cuidadosa consideración, podrían introducir cuellos de botella de información que afectarían el rendimiento.
- Swin Transformer también introdujo HTC++. Debido a que su implementación no es pública y a las restricciones en el tamaño de imagen del detector YoloR, no pudo ser implementado.

Arquitecturas

- Inspirados en la nomenclatura de YoloR, los modelos de YotoR se **nombran** utilizando la siguiente convención:

YotoR {Backbone Swin}{Cabeza YoloR}{# Bloques}

- Implementamos las siguientes arquitecturas:

YotoR TP5

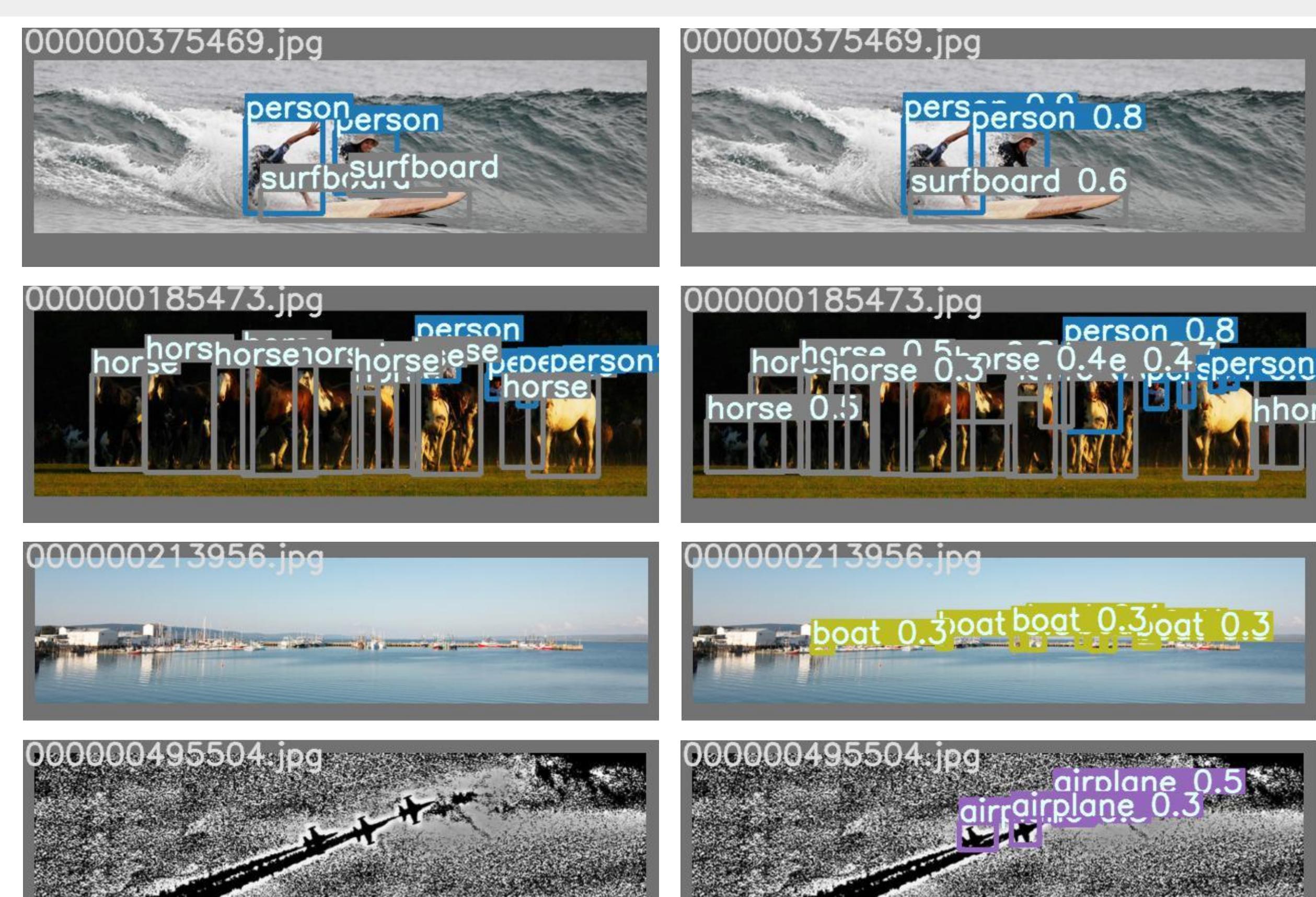
YotoR BP4

YotoR BB4

Resultados

- Los modelos YotoR, en comparación con sus backbones Swin con Cascade R-CNN, presentan **tiempos de inferencia menores**, con TP5 reduciendo un **50%**, y con BP4 y BB4 reduciendo en un **40%**.
- Los modelos TP5 y BP4 **superan a los baseline**, incluyendo YoloR P6, en términos de mean Average Precision (mAP) en val2017, mientras que BB4 no alcanza los resultados esperados.
- Los modelos YotoR **superan a sus backbones Swin tanto en rendimiento como en velocidad**, como se muestra en la relación entre el tiempo de inferencia y el mAP, lo que sugiere una mejora significativa.

Fig. 4 (Izquierda) Imágenes y anotaciones del set de test. (Derecha) Predicciones de YotoR BP4



Conclusiones

- Presentamos **YotoR**, una arquitectura híbrida que fusiona los modelos Swin Transformer y YoloR para la detección de objetos, demostrando **mejoras en precisión como en velocidad** de inferencia.
- Las implicaciones más amplias se extienden al diseño de modelos basados en transformers para tareas de imagen y arquitecturas de múltiples tareas.
- Las direcciones futuras de investigación incluyen la **actualización del backbone de YotoR**, investigar la integración de conocimiento implícito y la consideración de redes multimodales para mejorar el rendimiento en la detección de objetos.

Referencias

- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. Attention Is All You Need. (2017)
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J. & Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. (2020)
- Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S. & Guo, B. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. (2021)
- Wang, C., Yeh, I. & Liao, H. You Only Learn One Representation: Unified Network for Multiple Tasks. (2021)

Objetivo

- El objetivo de este trabajo es crear un **modelo híbrido** para la detección de objetos que aproveche el backbone del **Swin Transformer**, mientras se fusiona simultáneamente con el neck y la cabeza multipropósito de la arquitectura **YoloR**.

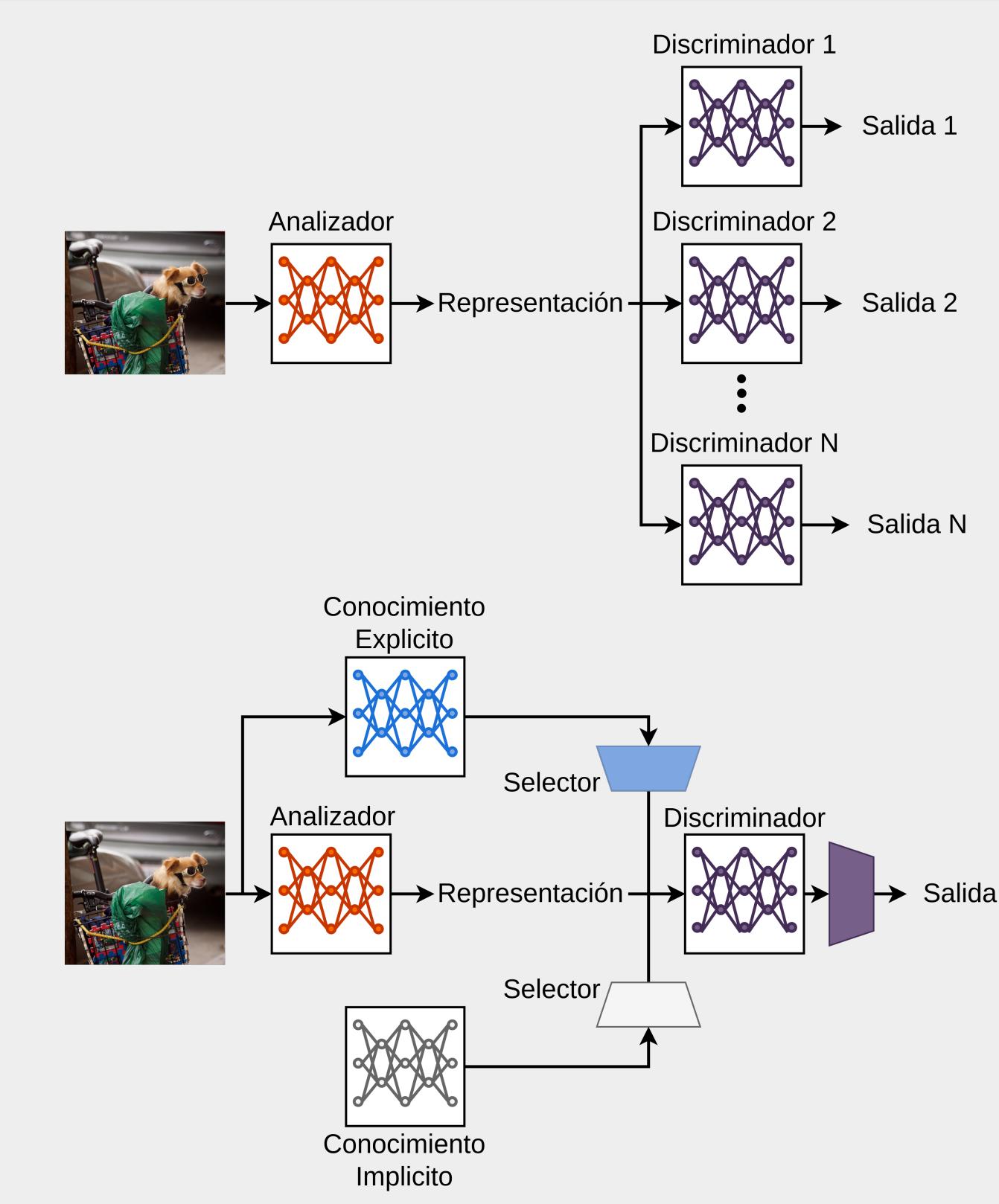


Fig. 2 Diferencia entre una arquitectura tradicional para múltiples tareas y una unificada con conocimiento implícito

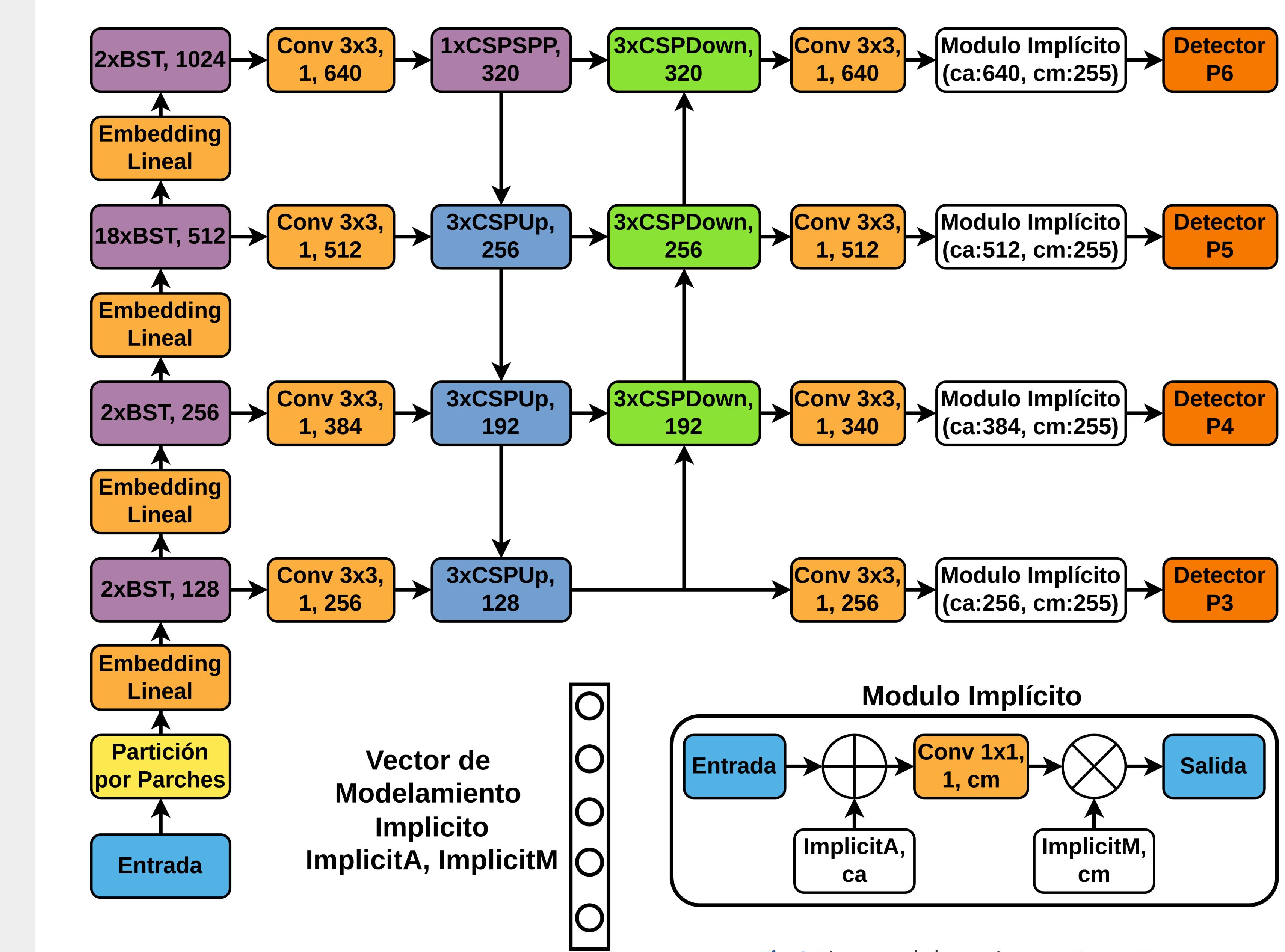


Fig. 3 Diagrama de la arquitectura YotoR BP4

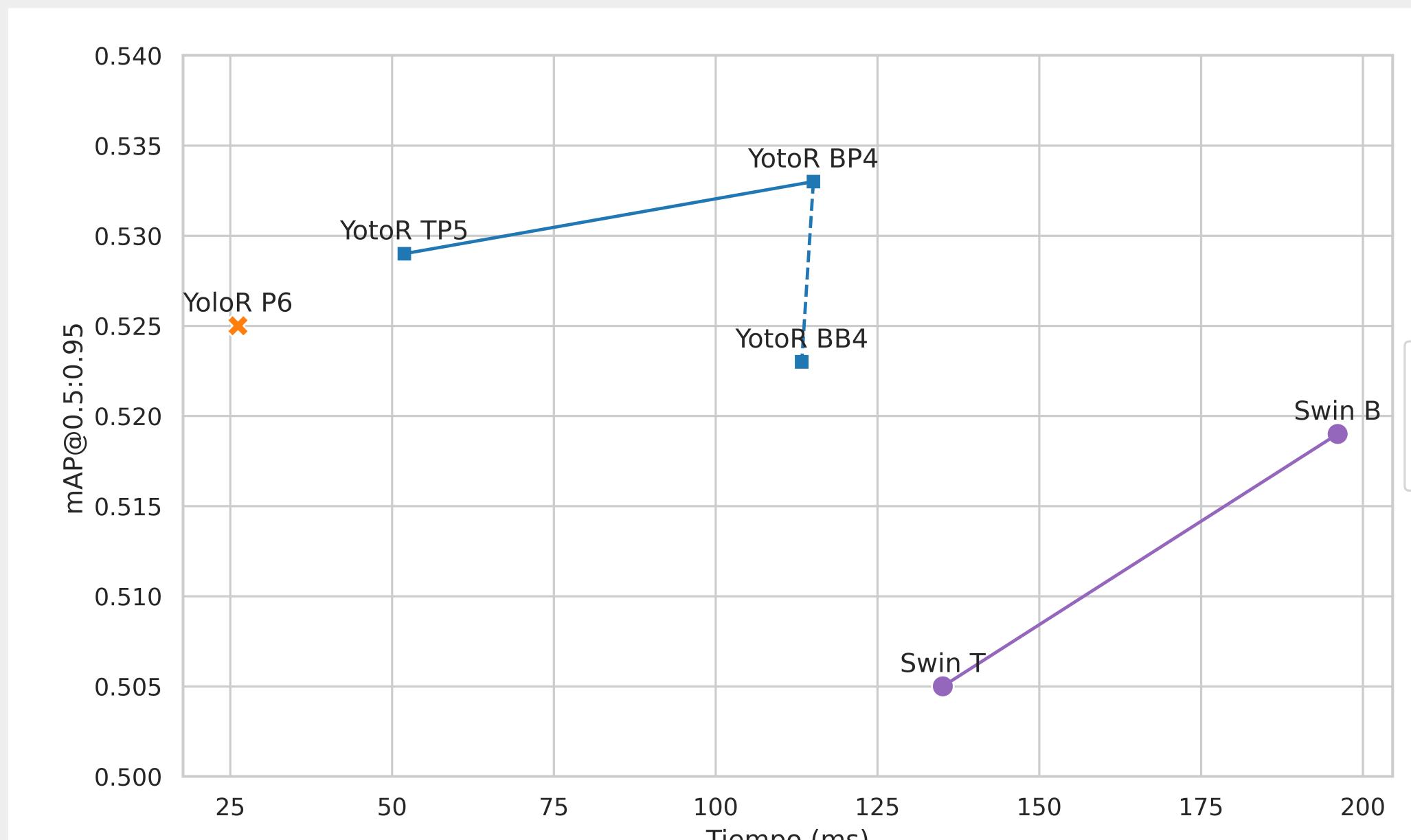


Fig. 5 Comparación de desempeño entre YoloR, Swin Transformer y YotoR