MixMatchDet: Mixing and Matching Tokens for Plain ViT Detectors

David Rohrschneider, Anselm Haselhoff, Uwe Handmann

Ruhr West University of Applied Sciences, 46236 Bottrop, Germany

1 Motivation

- DETR-Architectures are complex and detection is increasingly shifted towards the backbone
- Vision foundation models provide highly informative & dense feature representations
- Previous works show that relying solely on special [DET] Tokens leads to slow convergence and limited performance

How can useful object hypotheses be derived from encoded ViT representations?

2 Method

YOLOS Baseline

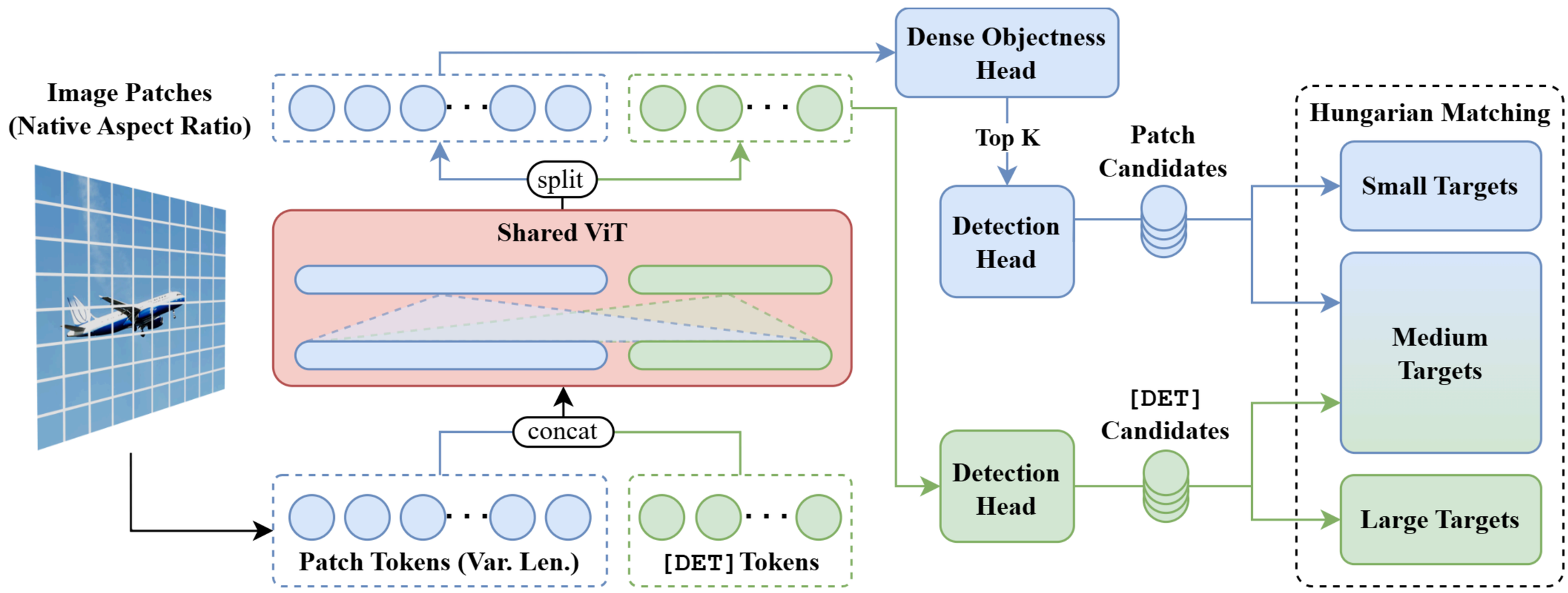
N_{DET} learnable [DET] tokens and patch tokens create the input sequence for DINOv2 and DINOv3 models

Mixing the candidates

Top-K patch tokens, selected via an objectness head, and all [DET] tokens form the unified candidate set

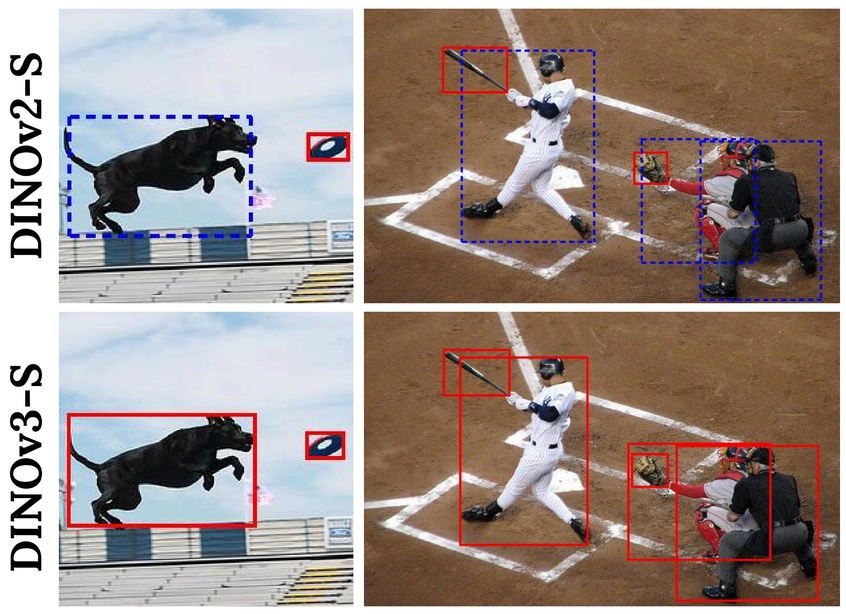
Matching the targets

To study the roles of each candidate type, we apply Scale-aware Matching based on target object size



3 Results (COCO Dataset)

Model	Epochs	AP	#Params (M)
YOLOS-S	150	36.6	24.6
DINOv2-S	50	44.1	24.2
DINOv2-B	20	46.4	93.1
DINOv2-L	20	51.1	314
DINOv3-S	50	44.6	22.5
DINOv3-B	20	50	88.9
DINOv3-L	20	55.2	308



Finding #1

Mixing candidate types substantially increases AP

Finding #2

[DET] tokens are global and match larger objects, patch tokens are spatially anchored and match smaller objects

Finding #3

Scale-aware matching helps DINOv2, while DINOv3 almost exclusively uses patch tokens

Kofinanziert von der Europäischen Union

Ministerium für Wirtschaft, Industrie, Klimaschutz und Energie des Landes Nordrhein-Westfalen

www.efre.nrw

prosperkolleg
Transferhub - Digitalisierung & Circular Economy

HRW HOCHSCHULE RUHR WEST
UNIVERSITY OF APPLIED SCIENCES

PROMOTIONS-KOLLEG NRW