Horizon 2020
European Union Funding
for Research & Innovation

Efficient Feature Extraction for Non-Maximum Suppression in Visual Person Detection

CHARALAMPOS SYMEONIDIS, IOANNIS MADEMLIS,
IOANNIS PITAS & NIKOS NIKOLAIDIS

ICASSP 2023

INTRODUCTION

- Non-Maximum Suppression (NMS) is a final refinement step incorporated to almost every visual object detection framework.
- NMS task is to prune the number of overlapping detected candidate Regions-of-Interest (RoIs) and replace them with a single and spatially accurate detection.
- Most NMS methods struggle to perform when they operate on images depicting objects in complex scenes, where several in-between occlusions appear.
  - This occurs frequently when detecting persons/pedestrians within human crowds.
- In this work, we propose FSeq^2-NMS which:
  - incorporates an *appearance-based RoI representations extraction module*, capable of utilizing feature maps precomputed by the intermediate layers of a detector.
    - The RoI representations can be used by a neural network architecture [1], suitable for discriminating duplicate RoIs in the challenging person detection task.
  - can be easily plugged on top of any DL-based detector and trained as a separate submodule.
  - outperforms SoA NMS methods on the challenging person detection task.

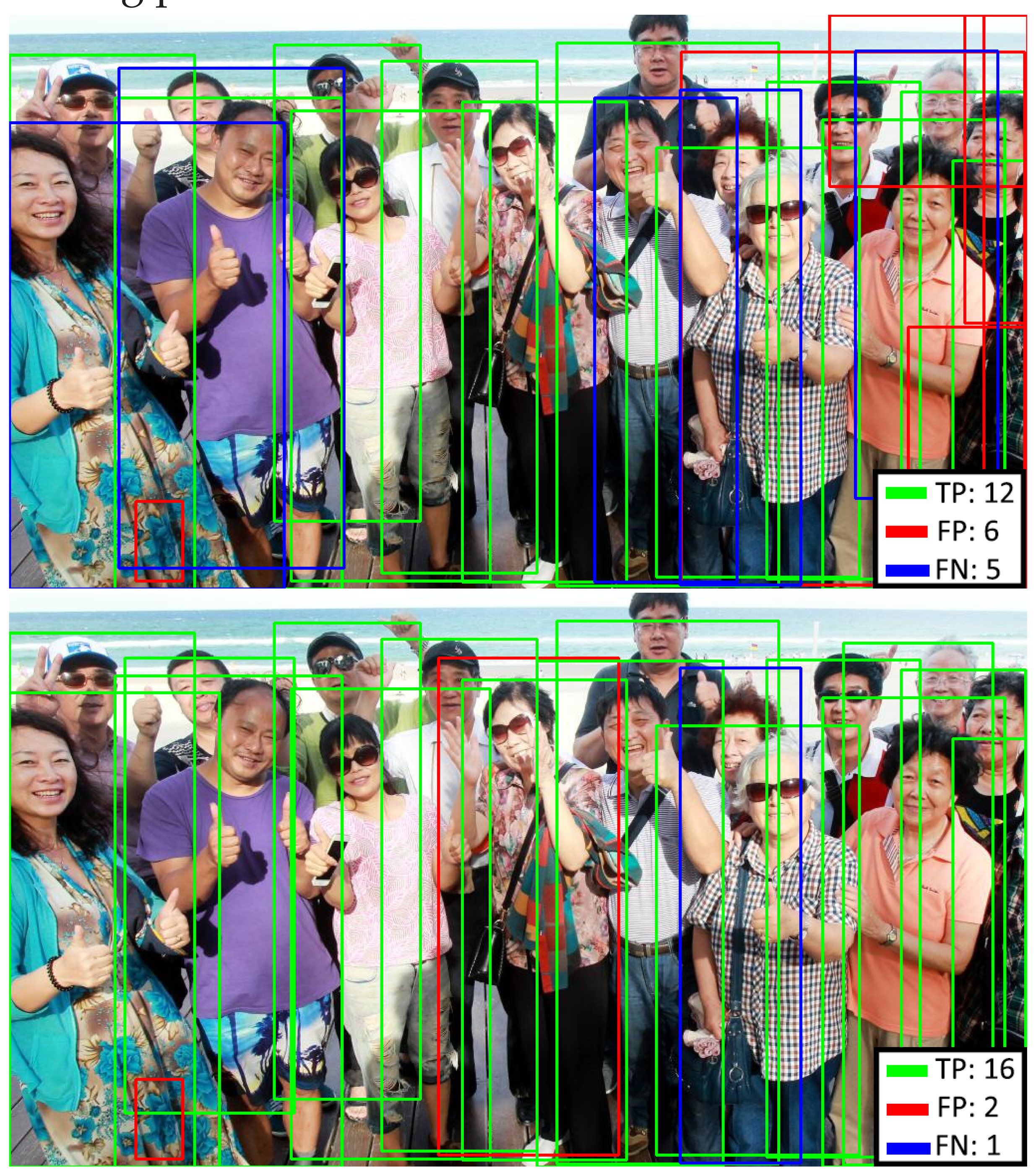

Figure 1: *Top:* Detections after applying GreedyNMS at 0.5 IoU.
Bottom: Detections after applying the proposed FSeq^2-NMS.

PIPELINE OF FSEQ^2-NMS

- The *Appearance-based RoI Representations Extraction Module* is incorporated in Seq2Seq-NMS [1].
  - The original architecture of Seq2Seq-NMS formulates NMS as a sequence-to-sequence problem, exploits the Multihead Scale-Dot Product Attention mechanism and jointly processes both geometric and visual properties of the input candidate RoIs.
  - It replaces the GPU-bound low-level Frame Moments Descriptor (FMoD) in the original variant of Seq2Seq-NMS.
- The proposed module should be trained along with core attention-based Seq2Seq-NMS.
- The training procedure of FSeq^2-NMS must be carried out after the training of the deployed detector.

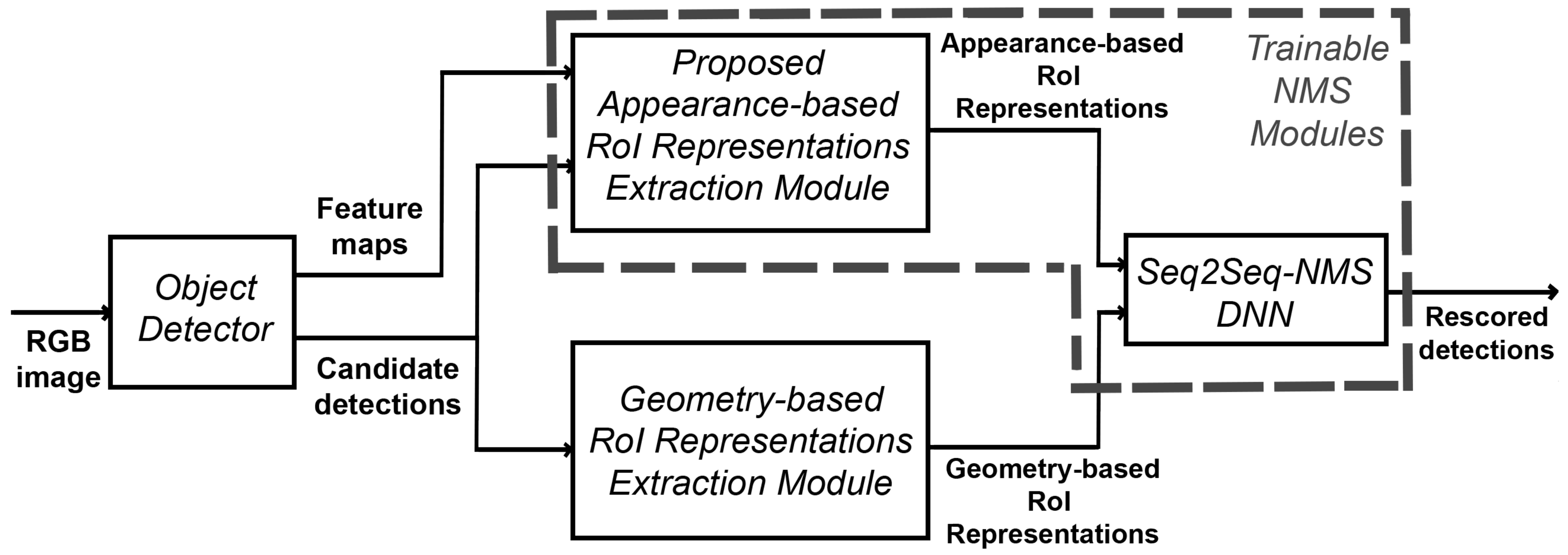

Figure 2: Pipeline of the object detection framework, in which FSeq^2-NMS in employed.

APPEARANCE-BASED ROI REPRESENTATIONS EXTRACTION MODULE

- As input the module receives:
  - the coordinates of N candidate RoIs.
  - a set of features maps, extracted from an in-between layer of the detector and resized to a fixed 64×64 resolution.
- RoI maps are in-parallel extracted from input feature-maps, using the RoIAlign [2] operator, in a fixed 20×20 spatial resolution.
- Two convolutional layers, followed by a max-pooling layer are applied on the extracted RoI maps.
- The final appearance-based RoI representations $\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, .., \mathbf{a}_N] \in \mathbb{R}^{N \times d_a}$ are computed by flattening the RoI maps and applying a fully connected layer.

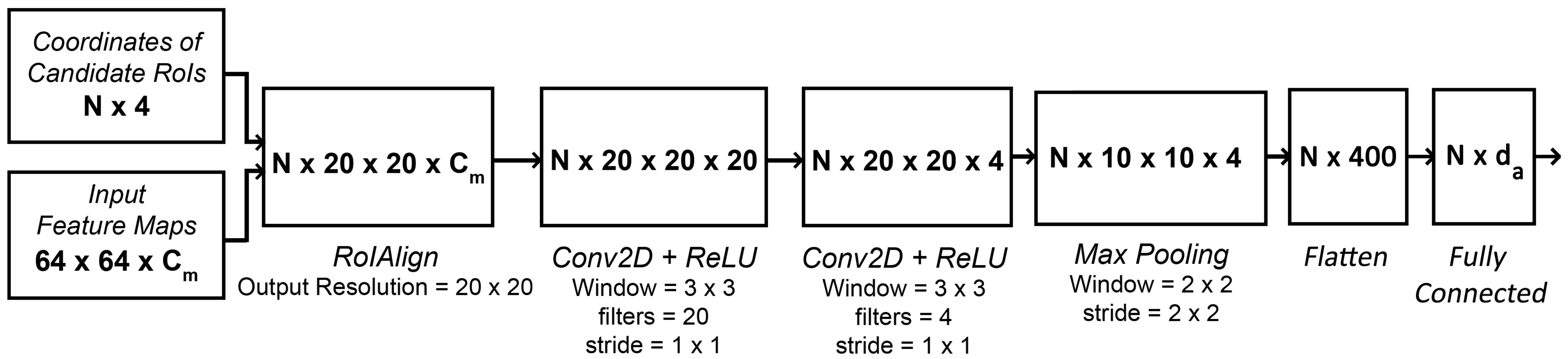

Figure 3: The proposed appearance-based RoI representations extraction module, capable to utilize feature-maps of the corresponding detector. C_m corresponds to the number of the channels of the input feature maps and d_a corresponds to the dimension of the final appearance-based RoI representations.

EXPERIMENTAL EVALUATION

- The performance of the FSeq^2-NMS was evaluated on PETS and CrowdHuman datasets.
  - Both datasets contain scenes depicting humans in crowded scenes.
- The Single Shot Detector (SSD) was selected.
  - The detector was trained from scratch.
  - VGG16 with atrous convolutions was selected as the backbone CNN.
  - Feature-maps from the initial layer of VGG16 were selected as input to the appearance-based RoI representations extraction module.

EXPERIMENTAL EVALUATION (CONTD.)

Table 1: Comparison of different NMS methods on PETS dataset.

Method	Device	Test set		Average Inference Time (ms)
		$AP_{0.5}$	$AP_{0.5}^{0.95}$	
Greedy-NMS	CPU	89.9%	36.3%	13.1
TorchVision NMS	GPU	90.0%	36.4%	0.2
Soft-NMS$_L$	CPU	90.0%	38.2%	108.8
Soft-NMS$_G$	CPU	89.6%	38.6%	134.4
Cluster-NMS	GPU	90.2%	36.9%	13.4
Cluster-NMS$_D$	GPU	90.2%	36.6%	17.9
Cluster-NMS$_{S+D}$	GPU	90.6%	38.3%	22.4
GossipNet	GPU	90.7%	38.8%	24.5
FSeq^2-NMS	GPU	**91.2%**	**38.9%**	7.8
Gains		+0.5%	+0.1%	-

Table 2: Comparison of different NMS methods on CrowdHuman dataset.

Method	Device	Test set		Average Inference Time (ms)
		$AP_{0.5}$	$AP_{0.5}^{0.95}$	
Greedy-NMS	CPU	67.0%	32.4%	9.8
TorchVision NMS	GPU	66.9%	32.4%	0.4
Soft-NMS$_L$	CPU	66.5%	32.3%	54.2
Soft-NMS$_G$	CPU	67.1%	33.0%	58.1
Cluster-NMS	GPU	67.1%	32.1%	5.0
Cluster-NMS$_D$	GPU	67.1%	32.1%	6.5
Cluster-NMS$_{S+D}$	GPU	65.7%	31.8%	8.0
GossipNet	GPU	72.4%	35.0%	10.0
FSeq^2-NMS	GPU	**75.3%**	**36.9%**	4.8
Gains		+2.9%	+1.9%	-

- The results confirm that exploiting semantic visual appearance descriptions of the candidate RoIs, along with their geometric interrelations, is the best option for NMS in the person detection task.
- The use of feature maps, extracted during the inference phase of the object detector, allows FSeq^2-NMS to achieve fast inference times on GPU, compared to most baseline methods.

REFERENCES

[1] C. Symeonidis, I. Mademlis, I. Pitas, and N. Nikolaidis. Neural attention-driven non-maximum suppression for person detection. *IEEE Transactions on Image Processing (TIP)*, 32:2454–2467, 2023.

[2] K. He, G. Gkioxari, P. Dollar, and R. Girshick. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

ACKNOWLEDGEMENTS

This research has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 871449 (OpenDR).