EFFICIENT FEATURE EXTRACTION FOR NON-MAXIMUM SUPPRESSION IN VISUAL PERSON DETECTION

Charalampos Symeonidis, Ioannis Mademlis, Ioannis Pitas and Nikos Nikolaidis

Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
E-mail: {charsyme, imademlis, pitas, nnik}@csd.auth.gr

ABSTRACT

Non-Maximum Suppression (NMS) is a post-processing step in almost every visual object detector, tasked with rapidly pruning the number of overlapping detected candidate rectangular Regions-of-Interest (RoIs) and replacing them with a single, more spatially accurate detection (in pixel coordinates). The common Greedy NMS algorithm suffers from drawbacks, due to the need for careful manual tuning. In visual person detection, most NMS methods typically suffer when analyzing crowded scenes with high levels of in-between occlusions. This paper proposes a modification on a deep neural architecture for NMS, suitable for such cases and capable of efficiently cooperating with recent neural object detectors. The method approaches the NMS problem as a rescoring task, aiming to ideally assign precisely one detection per object. The proposed modification exploits the extraction of RoI representations, semantically capturing the region's visual appearance, from information-rich feature maps computed by the detector's intermediate layers. Experimental evaluation on two common public person detection datasets shows improved accuracy against competing methods, with acceptable inference speed.

***Index Terms*—** Non-Maximum Suppression, Object Detection, Scaled-Dot Product Attention, Person Detection, Deep Neural Networks

1. INTRODUCTION

Non-Maximum Suppression (NMS) is a final refinement step incorporated to almost every visual object detection framework, where any detected rectangular Regions-of-Interest (RoIs, defined in pixel coordinates) that spatially overlap are merged/filtered. The problem it attempts to solve arises from the tendency of many detectors to output multiple, neighbouring candidate object RoIs for a single given visible object, due to their implicit sliding-window nature. NMS methods typically rescore the raw candidate detections/RoIs outputted by the detector, before thresholding these modified scores so that, ideally, only a single RoI is finally retained for each visible object.

The de facto standard in NMS for object detection is GreedyNMS [1]. It selects high-scoring detections and deletes less confident neighbours, since they most likely cover the same object. Its simplicity and speed make it competitive against proposed alternatives, given that rapid execution is of utmost importance in NMS. An Intersection-over-Union (IOU) threshold determines which less-confident neighbors are suppressed by a detection. Most NMS algorithms, including GreedyNMS, do not make any extra effort to jointly process the RoIs and assign one detection per object. In addition, this fixed IOU threshold leads GreedyNMS to failure in certain cases. For instance, wide suppression may remove detections that cover objects with lower scores, while too low a threshold is unable to suppress duplicate detections.

A typical case where most NMS methods struggle to perform is 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. This is a very important scenario for security- or safety-critical applications [2] [3] [4].

Over the years, several methods have been proposed as alternatives to GreedyNMS, achieving faster inference times or improved accuracy. Both non-neural algorithms and, more recently, deep learning (DL) have been employed to this end (see Section 2).

However, the vast majority of existing methods only exploit geometric properties/interrelations between the candidate RoIs, in the form of geometric features. An exception specifically for the case of person detection is Seq2Seq-NMS [5], a deep neural architecture which approaches NMS as a sequence-to-sequence problem. Seq2Seq-NMS extracts RoI representations based on geometric *and* visual appearance properties of the input candidate RoIs. An efficient implementation of FMoD [6] is employed for visual RoI description. These RoI representations are then refined by the Seq2Seq-NMS, by capturing relations of neighboring RoIs and aiming to ideally assign precisely one detection per person. A more recent variant of Seq2Seq-NMS, which used as input only geometric RoI representations, was presented in [7] showing that it can achieve improved accuracy rates when

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

deployed on distribution shift scenarios.

Motivated by the relative lack of NMS methods exploiting high-level, semantically meaningful representations of the candidate RoI/detections' visual appearance, this paper proposes a novel variant of Seq2Seq-NMS, named FSeq^2-NMS, which is able to harness the information-rich intermediate feature maps of DL-based object detectors. These are used to derive learned, high-level, semantically meaningful RoI representations, which are then exploited *instead of* handcrafted visual descriptors (such as [6]). FSeq^2-NMS can be easily plugged on top of any DL-based detector, and trained as a separate submodule. The efficacy achieved by the internal/latent image representations of state-of-the-art detectors allows the method to discriminate duplicate RoIs from a set of densely sampled and heavily occluded candidate detections, a problem commonly encountered when detecting humans in crowded scenes. Experiments conducted on two public person detection datasets, widely used for detecting humans in crowded scenes, confirms that FSeq^2-NMS is highly suitable for this scenario, achieving top accuracy.

2. RELATED WORK

Modern attempts to replace GreedyNMS and improve upon it were initially non-neural. Thus, Soft-NMS [8] employs a rescoring function aiming to decrease the score of neighboring less-confident detections, instead of completely eliminating them, achieving better precision and recall rates compared to GreedyNMS. Gaussian and linear weighting functions are utilized, which both require a hyper-parameter tuning similar to GreedyNMS. In [9], the authors replaced the classification scores of candidate detections, used in GreedyNMS, with learned localization confidences to guide NMS towards preserving more accurately localized bounding boxes.

A number of more advanced methods rely on *Distance-IoU* (DIoU) [10], a new metric which can replace the typical IoU metric in GreedyNMS. [10] suggested that the suppression procedure should take into account not only the overlap of two neighboring detections, but also the distances between their centers. Alternatively, Cluster-NMS was proposed in [11], i.e., a technique where NMS is performed by implicitly clustering candidate detections. Cluster-NMS can incorporate geometric factors to improve both precision and recall rates and can efficiently run on a GPU, achieving very fast inference runtimes. In [12], the authors presented Representative Region NMS, an approach to effectively remove the duplicate candidate detections in human crowded scenes. The method uses the IoU between the visible parts of two RoIs to determine whether the two full-body boxes are overlapped. In the pedestrian detection task, the novel attribute-aware MMS [13] was proposed, in order to distinguish the pedestrian from a high overlapped group. The proposed method adaptively rejects the false-positive results in the crowded settings.

Due to the prevalence of DL, neural NMS methods started to appear during the late 2010s. In [14], an attention module exploited relations between the input detections, in order to classify them as duplicate or not. Adaptive-NMS [15] is a dynamic thresholding version of GreedyNMS, specifically for detecting humans within crowds. A relatively shallow neural network was designed for predicting a density map and then the proposed method set an adaptive IoU thresholds in NMS for different detections according to the predicted density. In [16], the authors presented GossispNet, a DNN-based NMS method, which jointly analyzes the scores and coordinates of candidates detections in the image, so as not to directly prune them, but to rescore them. GossipNet was modified in [17], for the specific case of person detection from aerial views, by exploiting the handcrafted FMoD descriptor vectors [6] for representing the visual appearance of the candidate RoIs. Seq2Seq-NMS [5], upon which the method presented in this paper relies, also exploited FMoD descriptors for visual RoI representation and incorporated them into a sequence-to-sequence DL neural architecture for candidate RoI rescoring, operating via the Scaled-Dot Product Attention mechanism. Finally, NMS-Loss [18] can be incorporated to almost any single class DL-based object detector, allowing it to be trained with NMS end-to-end and pay attention to the false predictions caused by NMS.

3. EFFICIENT FEATURES FOR DEEP NEURAL NMS IN OBJECT DETECTION

Seq2Seq-NMS [5] is a DL-based NMS method, aiming to classify an input candidate detection as "correct" or as "potentially suppressed". The label of each candidate detection is formed based on evaluation criteria established in object detection [19] [20]. The method mainly relies on the Scaled Dot-Product Attention mechanism, for exploring relations between neighboring candidate RoIs and finally build discriminative RoI representations for the classification task. As input, Seq2Seq-NMS receives appearance-based and geometric representations for each candidate RoI. In the original version of the method, the authors proposed the use a fast and parallel GPU-bound implementation of FMoD [6], as an optional step, for extracting appearance-based RoI representations. In this implementation, FMoD computed an edge map of the corresponding image, and then extracted an appearance-based representation for each RoI based on statistical properties (e.g., mean, skew, etc.,) of the spatial distribution of the enclosed edges. Based on the corresponding experimental evaluation, Seq2Seq-NMS, along with FMoD, achieved top results against competing methods, proving to be a suitable solution for Non-Maximum Suppression on the person detection task.

Aiming to exploit the representational efficacy of modern DL-based detectors, *this paper proposes FSeq^2-NMS: an improvement of Seq2Seq-NMS, which incorporates an appearance-based RoI representations extraction module, capable of utilizing feature maps precomputed by the intermediate layers of the employed detector*. Thus, FSeq^2-NMS

outperforms the baseline Seq2Seq-NMS on the challenging task of detecting humans within crowded scenes, since it utilizes higher-quality, semantically meaningful representations of the raw candidate ROIs outputted by the object detector.

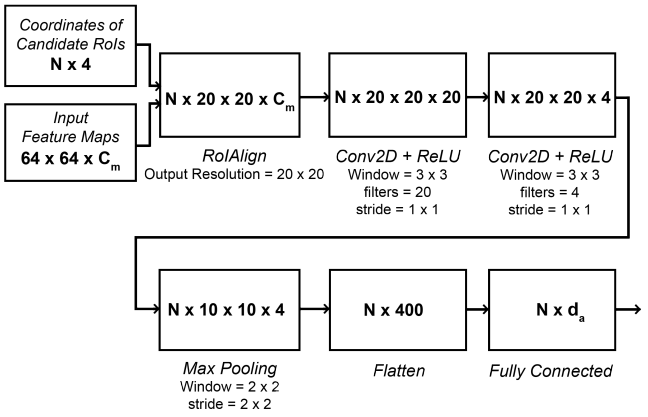

Fig. 1: Appearance-based RoI representations extraction module, capable to utilize feature-maps of the corresponding detector.

The proposed architecture of the appearance-based Roi representations extraction module is depicted on Fig. 1. As input, the module receives $\mathbf{B} = [\mathbf{b}_0, \mathbf{b}_1, .., \mathbf{b}_N] \in \mathbb{R}^{N\times 4}$, which correspond to the coordinates of N candidate RoIs, as well as $\mathbf{M} \in \mathbb{R}^{64\times 64\times C_m}$, which correspond to a set of features maps, extracted from an in-between layer of the deployed detector and resized to a fixed 64×64 resolution. C_m corresponds to the number of the channels of the corresponding feature maps. Using the *RoIAlign* [21] operator, initial RoI maps can be in-parallel extracted in a fixed 20×20 spatial resolution. Then two convolutional layers, with the Rectified Linear Unit (ReLU) as activation function, followed by a max-pooling layer are applied on the extracted RoI maps. The final 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 using ReLU as activation function. d_a corresponds to the dimension of the final appearance-based RoI representations.

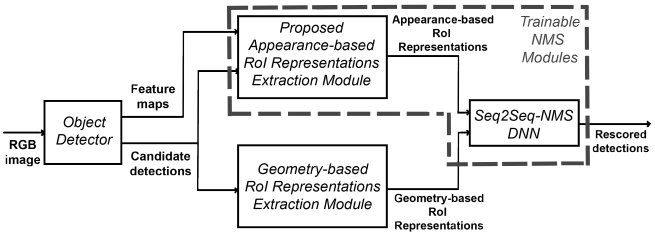

Fig. 2: Pipeline of the overall object detection framework, in which FSeq^2-NMS in employed.

The proposed appearance-based RoI representations extraction module should be trained along with core attention-based Seq2Seq-NMS, as it consists part of the overall structure. However, similar to [5], the training procedure of the proposed architecture of FSeq^2-NMS must be carried out after the training of the deployed detector. The overall detection pipeline, which incorporates FSeq^2-NMS, is depicted in Fig. 2.

4. EXPERIMENTAL EVALUATION

The performance of the proposed Seq2Seq-NMS variant was evaluated on two separate datasets, suitable for detecting humans in crowded scenes. In both datasets, candidate RoIs from the *Single Shot Detector* (SSD) [22] were provided as input to the corresponding NMS methods. In the implemented version of the detector, VGG16 with atrous convolutions was selected as the backbone CNN. The input images were resized to a resolution of 512×512 pixels, while the detector was trained from scratch for each dataset[1].

The core attention-based architecture of Seq2Seq-NMS and the training setup were similar to [5] and any deviations are reported separately on each dataset. Feature-maps from the initial layer of VGG16 were selected as input to the employed appearance-based RoI representations extraction module. Based on this selection, $C_m = 512$. In addition, we set $d_a = 315$.

In both datasets, FSeq^2-NMS was compared against neural and non-neural NMS algorithms. The first competing method is a baseline Greedy NMS approach running on GPU. The second is TorchVision's[2] GreedyNMS implemented to run very fast on GPUs. Soft-NMS [8], i.e., a non-neural NMS method widely used as a more accurate replacement for Greedy NMS, was also tested. Additionally, several variants of Cluster-NMS [11], a more recent non-neural method, were also used for comparisons. The last approach selected for comparison purposes is GossipNet [16], a neural NMS method achieving state-of-the-art accuracy. More details regarding these variations can be found in [11]. Additional information about the variant of each competing method can be found in [5].

The hyperparameters of all non-neural methods were tuned so as to report the best achieved results on 0.5 IoU matching threshold. Evaluation was performed on a PC using an Intel Core i7-7700 CPU and an NVIDIA GeForce RTX 2080 GPU with 11GB of memory, both for training and inference. The employed evaluation metrics are $\text{AP}_{0.5}$, $\text{AP}_{0.5}^{0.95}$ and inference times. $\text{AP}_{0.5}$ corresponds to the average precision for 0.5 IoU, while $\text{AP}_{0.5}^{0.95}$ to the mean average precision for IoU ranging from 0.5 to 0.95 with a step size of 0.05. Finally, all RoIs outputted by the NMS algorithms were utilized for evaluation, without any thresholding.

4.1. PETS

PETS [23] is a relatively small dataset, whose images were collected from static surveillance cameras and provide diverse

[1]The employed SSD implementation was adopted from `https://github.com/opendr-eu/opendr/tree/master/src/opendr/perception/object_detection_2d/ssd`

[2]`https://pytorch.org/vision/stable/ops.html#torchvision.ops.nms`

levels of occlusion.

The average number of people depicted in an image is approximately 14. The proposed NMS architecture was trained for 6 epochs. The learning rate was set to $10^{-4}/10^{-5}/10^{-6}$ for epochs 1-3/4-5/6, respectively. GossipNet's architecture and training followed [16]. Final parameters of all methods were selected according to the best achieved accuracy in the validation set.

Table 1 reports the results using cadidate detections from [22]. FSeq^2-NMS achieved an $AP_{0.5}$ of 87.4% and an $AP^{0.95}_{0.5}$ of 38.0% in the validation set, thus attaining gains of $+0.4\%$ $+1.0\%$ over GossipNet in the corresponding metrics. In the testing set, the proposed method achieved an $AP_{0.5}$ of 91.2% and an $AP^{0.95}_{0.5}$ of 38.9% surpassing GossipNet by $+0.5\%$ and $+0.1\%$ in the corresponding metrics.

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

Method	**Device**	**Val set**		**Test set**		**Average Inference Time (ms)**
		$AP_{0.5}$	**$AP^{0.95}_{0.5}$**	**$AP_{0.5}$**	**$AP^{0.95}_{0.5}$**	
Greedy-NMS	CPU	84.3%	34.7%	89.9%	36.3%	13.1
TorchVision NMS	GPU	84.3%	34.7%	90.0%	36.4%	0.2
Soft-NMS_L	CPU	85.3%	35.9%	90.0%	38.2%	108.8
Soft-NMS_G	CPU	83.9%	36.2%	89.6%	38.6%	134.4
Cluster-NMS	GPU	84.5%	34.3%	90.2%	36.9%	13.4
Cluster-NMS_S	GPU	84.7%	36.0%	90.1%	38.0%	13.8
Cluster-NMS_D	GPU	84.5%	34.5%	90.2%	36.6%	17.9
Cluster-NMS_{S+D}	GPU	85.7%	36.0%	90.6%	38.3%	22.4
Cluster-NMS_{S+D+W}	GPU	85.7%	36.0%	90.6%	38.3%	38.2
GossipNet	GPU	87.1%	37.0%	90.7%	38.8%	24.5
FSeq^2-NMS	GPU	**87.4%**	**38.0%**	**91.2%**	**38.9%**	7.8
Gains		+0.3%	+1.0%	+0.5%	+0.1%	-

4.2. CrowdHuman

The CrowdHuman dataset has been released specifically to target human detection in crowded areas, and has been proved to be a challenging for person detectors, due to heavy visual occlusions of individual humans. The average number of persons in an image is 22.64.

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

Method	**Device**	**Test set**		**Average Inference Time (ms)**
		$AP_{0.5}$	**$AP^{0.95}_{0.5}$**	
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_S	GPU	64.0%	31.0%	5.2
Cluster-NMS_D	GPU	67.1%	32.1%	6.5
Cluster-NMS_{S+D}	GPU	65.7%	31.8%	8.0
Cluster-NMS_{S+D+W}	GPU	65.7%	31.9%	32.3
GossipNet	GPU	72.4%	35.0%	10.0
FSeq^2-NMS	GPU	**75.3%**	**36.9%**	4.8
Gains		+2.9%	+1.9%	-

FSeq^2-NMS was trained for 14 epochs. The learning rate was set to $10^{-4}/10^{-5}/10^{-6}$ for epochs 1-8/9-12/13-14, respectively. GossipNet was trained for 10^6 iterations, with a learning rate set to 10^{-4} and decreased by 0.1 at the 6×10^5-th and the 8×10^5-th iterations.

Table 2 shows that the proposed method achieves gains, both in terms of $AP_{0.5}$ and $AP^{0.95}_{0.5}$. FSeq^2-NMS, achieved an $AP_{0.5}$ of 75.3% and an $AP^{0.95}_{0.5}$ of 36.9%, which are $+2.9\%$ and $+1.9\%$ improvements against GossipNet, respectively.

4.3. Discussion

FSeq^2-NMS achieved top accuracy rates on $AP_{0.5}$ and $AP^{0.95}_{0.5}$ metrics in both datasets, due to the incorporation of appearance-based RoI representations. These results demonstrate that the proposed module can push the core attention-based Seq2Seq-NMS DNN towards achieving top results in the challenging task of detecting humans in crowded scenes. This is done by exploiting the representational efficacy of modern-DL based detectors towards providing FSeq^2-NMS with enriched RoI representations regarding their visual appearance. Finally, the use of precomputed feature maps, already extracted during the inference phase of the object detector, allows the proposed NMS method to achieve very fast inference times on GPU, compared to most baseline methods.

5. CONCLUSIONS

Successful NMS is challenging when detecting humans in crowded areas with high levels of in-between occlusions. This paper proposed a modification to a DL-based NMS architecture, capable of harnessing the representational efficacy of state-of-the-art neural detectors. The proposed approach, called FSeq^2-NMS, is able to utilize feature maps, extracted from the intermediate detector layers, in order to build semantically rich representations of the candidate RoIs' visual appearance. These are then employed by the NMS Deep Neural Network which this paper improves (Seq2Seq-NMS), for better discriminating whether a candidate detection is duplicate or not. Experiments on two person detection datasets, whose images mostly depict humans in crowded scenes, showed that the proposed method is indeed suitable for such a scenario, achieving top accuracy rates among the competing methods. The results confirm that exploiting semantic visual appearance descriptions of the candidate RoIs is indeed the best option for NMS in person detection within dense crowd images, compared either to geometry-only RoI representations, or to using lower-level statistical visual appearance descriptors (e.g., FMoD).

6. REFERENCES

[1] P. F. Felzenszwalb, R. B. Girshick, D. A. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, pp. 1627–1645, 2009.

[2] C. Papaioannidis, I. Mademlis, and I. Pitas, "Fast CNN-based single-person 2D human pose estimation for autonomous systems," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[3] E. Kakaletsis, C. Symeonidis, M. Tzelepi, I. Mademlis, A. Tefas, N. Nikolaidis, and I. Pitas, "Computer vision for autonomous UAV flight safety: An overview and a vision-based safe landing pipeline example," *ACM Computing Surveys (CSUR)*, vol. 54, no. 9, pp. 1–37, 2021.

[4] E. Kakaletsis, I. Mademlis, N. Nikolaidis, and I. Pitas, "Multiview vision-based human crowd localization for UAV fleet flight safety," *Signal Processing: Image Communication*, vol. 99, pp. 116484, 2021.

[5] C. Symeonidis, I. Mademlis, I. Pitas, and N. Nikolaidis, "Neural attention-driven non-maximum suppression for person detection," *TechRxiv*, vol. 10.36227/techrxiv.16940275.v1, 2021.

[6] I. Mademlis, N. Nikolaidis, and I. Pitas, "Stereoscopic video description for key-frame extraction in movie summarization," in *Proceedings of the EURASIP European Signal Processing Conference (EUSIPCO)*, 2015.

[7] C. Symeonidis, I. Mademlis, I. Pitas, and N. Nikolaidis, "AUTH-Persons: A dataset for detecting humans in crowds from aerial views," in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, 2022, pp. 596–600.

[8] N. Bodla, B. Singh, R. Chellappa, and L. S. Davis, "Soft-NMS: Improving object detection with one line of code," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[9] B. Jiang, R. Luo, J. Mao, T. Xiao, and Y. Jiang, "Acquisition of localization confidence for accurate object detection," in *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, Springer.

[10] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," *Association for the Advancement of Artificial Intelligence (AAAI)*, vol. 34, no. 07, pp. 12993–13000, 2020.

[11] Z. Zheng, P. Wang, W. Ren, D.and Liu, R. Ye, Q. Hu, and W. Zuo, "Enhancing geometric factors in model learning and inference for object detection and instance segmentation," *IEEE Transactions on Cybernetics*, vol. 52, no. 8, pp. 8574–8586, 2022.

[12] X. Huang, Z. Ge, Z. Jie, and O. Yoshie, "Nms by representative region: Towards crowded pedestrian detection by proposal pairing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 10747–10756.

[13] J. Zhang, L. Lin, J. Zhu, Y. Li, Y.-C. Chen, Y. Hu, and S. C. H. Hoi, "Attribute-aware pedestrian detection in a crowd," *IEEE Transactions on Multimedia*, vol. 23, pp. 3085–3097, 2021.

[14] H. Hu, J. Gu, Z. Zhang, J. Dai, and Y. Wei, "Relation networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[15] S. Liu, D. Huang, and Y. Wang, "Adaptive NMS: Refining pedestrian detection in a crowd," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] J. Hosang, R. Benenson, and B. Schiele, "Learning Non-Maximum Suppression," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[17] C. Symeonidis, I. Mademlis, N. Nikolaidis, and I. Pitas, "Improving neural Non-Maximum Suppression for object detection by exploiting interest-point detectors," in *Proceedings of the IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, 2019.

[18] Z. Luo, Z. Fang, S. Zheng, Y. Wang, and Y. Fu, "NMS-Loss: Learning with non-maximum suppression for crowded pedestrian detection," in *Proceedings of the 2021 International Conference on Multimedia Retrieval*, 2021, p. 481–485.

[19] T.-Y. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2014.

[20] M. Everingham, L. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (VOC) challenge," *International Journal of Computer Vision*, vol. 88, pp. 303–338, 2009.

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

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, and A. Berg, "SSD: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[23] J. M. Ferryman and A. Ellis, "PETS2010: Dataset and challenge," in *Proceedings of the IEEE International Conference on Advanced Video and Signal-Based Surveillance (AVSS)*, 2010.