MULTI-SCALE DEFORMABLE TRANSFORMER ENCODER BASED SINGLE-STAGE PEDESTRIAN DETECTION

Jing Yuan[1], *Panagiotis Barmpoutis*[2], *Tania Stathaki*[1]

[1] Department of Electronic and Electronic Engineering, Imperial College London, London, UK
[2] Department of Computer Science, University College London, London, UK

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897361

ABSTRACT

Pedestrian detection is a key task in intelligent video surveillance systems which requires both fast inference and high detection accuracy. Although single-stage deep learning pedestrian detectors have achieved relatively high detection accuracy with simpler architecture and less inference time, their performance is limited compared to two-stage methods. The reason is the lack of scale-aware features without the assistance of proposal regions. To overcome this, a multi-scale deformable transformer encoder-based module is proposed. It can extract the sparse important features at deformable sampling locations from multiple levels. The proposed architecture significantly improves the performance compared to the baseline center and scale prediction method on both Caltech and Citypersons datasets. It even outperforms the state-of-the-art two-stage methods in detecting heavily occluded pedestrians on Citypersons validation set.

Index Terms— Pedestrian detection, single-stage method, vision transformer

1. INTRODUCTION

Deep learning techniques have been applied to pedestrian detection due to their automatic and effective feature extraction compared to traditional handcrafted features like HOG [1] and ACF [2]. These detectors can be divided into two-stage and single-stage methods. The former is easier to achieve leading performance on various challenging pedestrian benchmarks such as Caltech [3] and Citypersons [4] datasets. Two-stage detectors [5-7] make the second time prediction using scale-aware features extracted in Region Of Interests (ROIs). Such a strategy refines the detection but is time-consuming. To this end, single-stage methods are proposed to achieve a better trade-off between accuracy and inference time. They discard ROI and only predict once in each inference. However, the lack of ROI results in fixed size of the regions from which the final features are extracted. This means that such features fed to the detection head may either contain too much background information for relatively small and occluded objects or the information is not comprehensive enough for large objects. Therefore, it is still difficult for single-stage methods to outperform two-stage methods especially in detecting occluded pedestrians.

To overcome such intrinsic disadvantage, we propose a module that selectively extracts useful features at deformable locations. The module is based on a single multi-scale deformable transformer encoder inspired by the recently proposed deformable DEtection TRansformer (DETR) [8]. Initially, the sparse sampling locations spread over the global feature map with uniformly initialized weights. After the training, the module looks at all sparse features and pays more attention to only important ones and suppresses the background information. The proposed module is adaptive to pedestrians with varying scales and occlusion without any assistance of ROI, thus making the best of the fast inference and scale-aware features. We build our detector upon Center and Scale Prediction (CSP) protocol [9] and choose it as the baseline in the following comparisons. This paper makes the following contributions:

- A novel multi-scale deformable feature extractor is proposed for single-stage pedestrian detection.
- The performance gap between single-stage and two-stage methods is narrowed.
- The application of vision transformers in pedestrian detection tasks is further extended.

2. RELATED WORKS

2.1. Single-stage Pedestrian Detection

Single-stage methods [9-12] accomplish localization and classification within a single stage. They work faster than two-stage methods, and they heavily rely on feature fusion. Among them, CSP is one of the first anchor-free single-stage pedestrian detectors. It concatenates multi-level feature maps and predicts a center heatmap, a scale map and an optional offset map. Each pixel in these three maps represents a confidence score, the corresponding logarithm height, and the offsets of the center position in x and y directions respectively. Most single-stage methods focus on the design of multi-level features [13-17]. For example, [14, 17] established a feature pyramid network (FPN) similar architecture but modulates multi-scale feature maps with channel and spatial attention weights before up-sampling.

2.2. Vision Transformer

The vision transformer is an application of the transformer originating from Natural Language Processing (NLP) to Computer Vision (CV). It can be used uniquely or jointly with convolutional blocks, for example, ViT [18] establishes a backbone with pure attention models while DETR [19] applies encoder-decoder vision transformers on top of the convolutional backbone. Our architecture falls in the latter category. Vision transformers apply an attention mechanism to quantify pairwise long range entity interactions [20]. Deformable DETR [8], a variant of DETR, models the relations between global features with multi-scale deformable self-attention. It is a fast-converging and memory-saving vision transformer which facilitates multi-scale feature maps with high resolutions. It outperforms Faster R-CNN [5] and DETR on COCO 2017 validation set [21]. The potential of a vision transformer is appealing, but it is rarely applied in pedestrian detection. M. Lin et al. [22] discovered that deformable DETR performed worse than baseline Faster R-CNN on CrowdHuman dataset [23] and the training time increased tenfold. Although [22] closed the gap between (deformable) DETR and Faster R-CNN via the decoder with dense queries and rectified attention field, a current end-to-end pedestrian detector is not showing a significant advantage over traditional Faster R-CNN. To this end, we decide to adopt the deformable DETR encoder as a feature encoding module and apply it to single-stage methods to discover any possible improvements rather than putting the whole encoder-decoder based vision transformer into the pedestrian detector.

3. PROPOSED METHOD

3.1. Deformable Transformer Encoder

The deformable transformer encoder as shown in Fig. 1a takes multi-scale feature maps $\{\boldsymbol{x}^l\}_{l=1}^{L}$ ($L = 3$ is the number of feature maps) with height H_l and width W_l as inputs and outputs them with the same height and width. These feature maps together serve as the source feature map $\boldsymbol{z_s}$. The source feature map is embedded with fixed positional encodings and randomly initialized scale-level information to generate the query feature map $\boldsymbol{z_q}$. These feature maps and pre-generated reference points are fed into the multi-scale deformable attention module (Fig. 1b) to extract the deformable attention feature. Then it is added back to the source feature followed by a feed-forward network (Fig. 1c).

3.2. Multi-scale Deformable Attention Module

In Multi-Scale Deformable Attention Module (MSDAM), three multi-dimensional tensors, namely the value, weight, and location, are computed to produce the deformable attention feature map $\boldsymbol{z_o}$ via weighted average. As shown in Fig. 1(b), the source feature map $\boldsymbol{z_s}$ is encoded by a linear

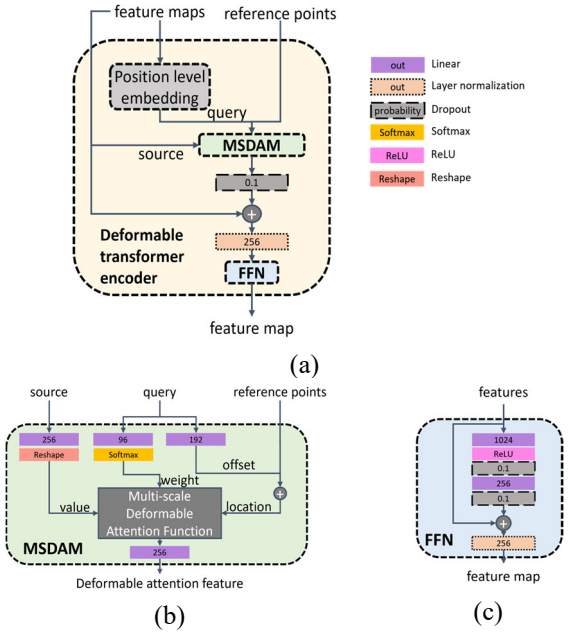

Fig. 1. A deformable transformer encoder (a) contains a Multi-Scale Deformable Attention Module (MSDAM) (b), a Feed-Forward Network (FFN) (c), and intermediate embedding and normalization layers.

layer to form the value tensor $\boldsymbol{v}$. The query feature map $\boldsymbol{z_q}$ is first fed into two linear layers respectively to predict the deformable attention weight $\boldsymbol{W}$ and sampling offset $\Delta\boldsymbol{p}$. The weight is further normalized by a softmax operator along the scale and sampling point dimensions. The sampling offset $\Delta\boldsymbol{p}$ is added to the reference points $\boldsymbol{p}$ to form the sampling locations. Then the value $\boldsymbol{v}$, deformable attention weight $\boldsymbol{W}$ and sampling location are sent to the multi-scale deformable attention function block to form the separate deformable attention feature $\boldsymbol{z}' \in \mathbb{R}^{N_q \times c_v}$ ($N_q = \sum_{l=1}^{L} H_l W_l$) for each attention head. To be specific, the q-th element of $\boldsymbol{z}'$ is expressed as

$$\boldsymbol{z}'_q = \sum_{p}^{N_p} \sum_{l=1}^{L} W_{plhq} \boldsymbol{v}_{\boldsymbol{p}_{ql} + \Delta\boldsymbol{p}_{qhlp}} \quad (1)$$

where q, h, l and p index the elements of the deformable attention feature $\boldsymbol{z}_o$, the attention head, the scale of value $\boldsymbol{v}$ and the sampling offsets. W_{plhq} is a value from the weight $\boldsymbol{W} \in \mathbb{R}^{N_q \times N_h \times L \times N_p}$. $\boldsymbol{p}_{ql}$ and $\Delta\boldsymbol{p}_{qhlp}$ denote the position of a reference point and the corresponding sampling offset from $\boldsymbol{p} \in \mathbb{R}^{N_q \times L \times 2}$ and $\Delta\boldsymbol{p} \in \mathbb{R}^{N_q \times N_h \times L \times N_p \times 2}$ respectively. The parameter $N_h = 8$ is the number of attention heads. The separate deformable attention features from N_h attention heads are projected to the q-th element of the final output deformable attention feature $\boldsymbol{z}_o$ by a linear layer expressed as

$$\boldsymbol{z}_{o_q} = \sum_{h=1}^{N_h} \boldsymbol{W}'_h \boldsymbol{z}'_{qh} \quad (2)$$

where $\boldsymbol{W}'_h \in \mathbb{R}^{c \times c_v}$ denotes the learnable weight for the h-th attention head. The vector $\boldsymbol{z}'_{qh} \in \mathbb{R}^{c_v}$ refers to the element $\boldsymbol{z}'_q$ obtained at h-th attention head.

3.4. Proposed Architecture

The proposed architecture consists of the ResNet 50 backbone, the deformable transformer encoder-based detection neck and the detection head as shown in Fig. 2. In the neck, feature maps C_1 and C_2 marked in Fig. 2 are upsampled by two times while C_3 is encoded with a convolutional layer. Three feature maps are fed into group normalization layers in case of the internal covariate shift (ICS) after the subsequent linear operations in the deformable transformer encoder. Subsequently, the multi-scale feature maps $\{\boldsymbol{x}^l\}_{l=1}^3$ are fed to the deformable transformer encoder to yield multi-scale feature maps. To simplify the network, only the smallest one (H/16, W/16) is sent to the detection head while the other two are discarded. Furthermore, in the head, the feature map is first upsampled by 4 times with a deconvolution layer for accurate pedestrian locating. In comparison to the architecture of CSP, the proposed detector adopts the novel deformable transformer encoder-based detection neck. The other parts of the proposed detector are similar to CSP however, a deconvolution layer is applied before the classifier and regressors in the detection head.

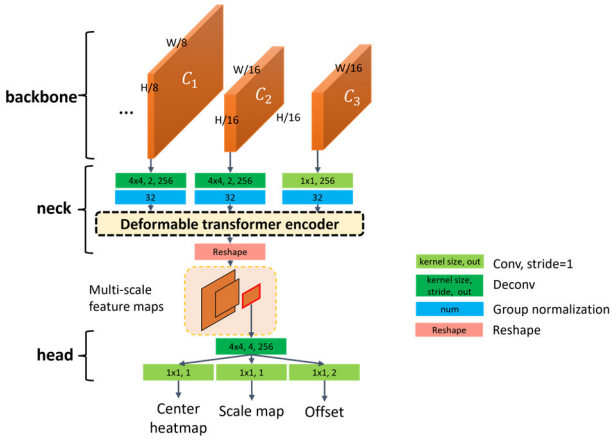

Fig. 2. The proposed network utilizes ResNet 50 as the backbone, followed by the deformable transformer encoder-based detection neck to produce multi-scale features, among which the smallest feature map is fed into the detection head.

3.5. Training and Inference

In training, to have a fair comparison, the same ground truth and loss functions as [9] are utilized. The center loss is formulated as:

$$L_{center} = -\frac{1}{K}\sum_{i=1}^{\frac{W}{4}}\sum_{j=1}^{\frac{H}{4}}\alpha_{ij}\left(1-\hat{p}_{ij}\right)^{\gamma}\log\left(\hat{p}_{ij}\right), \quad (3)$$

where

$$\hat{p}_{ij} = \begin{cases} p_{ij}, & if\ y_{ij}=1 \\ 1-p_{ij}, & \text{else} \end{cases} \quad (4)$$

$$\alpha_{ij} = \begin{cases} 1, & if\ y_{ij}=1 \\ \left(1-M_{ij}\right)^{\beta}, & \text{else} \end{cases}. \quad (5)$$

In (4) and (5), $p_{ij} \in (0,1)$ denotes the predicted center score and $\boldsymbol{M}$ is a Gaussian mask map with variances proportional to the height and width of each one of the K pedestrians. Based on experimental tests, the two hyper-parameters β and γ are set as 4 and 2 [9]. The scale and offset losses are calculated via smooth L1 loss:

$$L_{scale} = \frac{1}{K}\sum_{k=1}^{K} SmoothL1(s_k, t_k), \quad (6)$$

$$L_{offset} = \frac{1}{K}\sum_{k=1}^{K} SmoothL1(o_k, ot_k) \quad (7)$$

where s_k and o_k are the predicted height and offsets at k-th pedestrian with the ground truth t_k and ot_k. The overall loss function is

$$L = \lambda_c L_{center} + \lambda_s L_{scale} + \lambda_o L_{offset} \quad (8)$$

where λ_c, λ_s and λ_o are set as 0.01, 1 and 0.1.

In inference, the predicted centers with scores larger than the threshold (0.1 for Citypersons, 0.01 for Caltech) are kept and fed to NMS. The width of the bounding boxes is the multiplication of predicted scales and the constant 0.41.

4. EXPERIMENTS

4.1. Settings

The proposed network is trained and tested on the Caltech pedestrian dataset and Citypersons dataset. Note that for the Caltech dataset, the training data augmented by 10 folds containing 42782 images and the standard test set (4024 frames) with corresponding new annotations [24] are used.

We build our network on the reproduction of CSP provided by Pedestron [25]. The backbone is pretrained on ImageNet with Adam and moving average weights [26]. Standard data augmentation techniques including random horizontal flip, scaling and crop are applied. For the Caltech training set, random color distortion is also implemented. The input images are first resized to 336x448 and 640x1280 pixels for Caltech and Citypersons dataset. The proposed network is trained with a single NVIDIA GeForce RTX 3090 GPU for 9 and 75 epochs with batch size 16 and 4 on Caltech and Citypersons respectively. For the Caltech dataset, the base learning rate is 1e-4, which is decreased to 2e-5 after 5 epochs. For Citypersons, the base learning rate 2e-4 is decreased to 8e-5 after 40 epochs. All the training is performed with a randomly chosen and fixed seed.

Log-Average Miss Rate (denoted as MR^{-2}) over False Positive Per Image (FPPI) in the range $[10^{-2},10^{0}]$ is calculated over Reasonable (R) and Heavily Occluded (HO) subsets. R subset is the collection of pedestrians with a height larger than 50 pixels and visibility larger than 0.65. For the HO subset, the visibility lies in the range [0.2, 0.65].

4.2. Comparison with the State-of-the-arts

Extensive comparisons with the state-of-the-art pedestrian detectors on Caltech and Citypersons datasets are performed as shown in Table 1 and Table 2. Our method achieves the best performance among the listed detectors. It outperforms

the baseline CSP method [9] in both subsets on Caltech by 0.4% and 1.1% respectively. For Citypersons, the network converges quickly after about 56 epochs: MR^{-2} less than 11% and 40% with recall around 95.5% and 82.5% for R and HO subsets respectively. As shown in Table 2, although single-stage methods (ALFNet [12], PRNet [10] and CSP [9]) have achieved relatively competitive MR^{-2} in the R subset, two-stage methods like MGAN+ [27] and KGSNet [28] are showing overwhelming superiority in the HO subset with MR^{-2} less than 40%. However, our multi-scale deformable transformer encoder-based detector fills the gap between the single- and two-stage methods by reducing the MR^{-2} to 10.6% in the R subset, and 36.7% in the HO subset, which even outperforms the results of the two-stage methods by 3%.

4.3. Discussion

Ablation studies are conducted using Caltech and Citypersons datasets to verify the effectiveness of the proposed module. As presented in Table 3, using the proposed module reduces the MR^{-2} considerably in both R and HO subsets. More significant performance improvements in detecting heavily occluded pedestrians are observed. Preliminary experiments show that pedestrians under various scenarios can be detected, including some seriously occluded pedestrians which suffer from heavy inter or intraclass occlusion as presented in Fig. 3b-c.

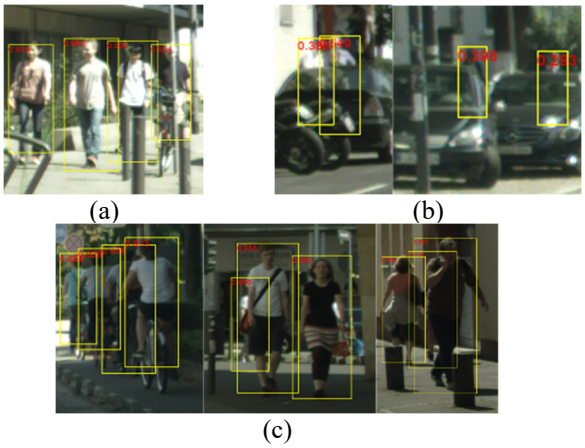

Fig. 3. Samples of detected pedestrians belonging to R subset (a) and heavily occluded pedestrians (b, c) from Citypersons validation set. Both pedestrians with inter class occlusion, such as cars (b), and intra class occlusion (c) are detected.

Our detector creatively adopts the vision transformer in pedestrian detection tasks and a novel scale-aware feature extraction module is proposed to facilitate single-stage detectors. This module adaptively sums the multi-scale multi-head features at deformable locations with learned weights, improving the overall performance. Experiments show that the detector is robust in various cases. Furthermore, it is particularly competitive in detecting occluded pedestrians even compared with two-stage methods on the difficult Citypersons dataset.

Table 1. Comparison with the sate-of-the-art pedestrian detectors on Caltech test set in terms of MR^{-2} (%).

Method	R	HO
Faster R-CNN [5]	8.7	53.1
RPN+BF [29]	7.3	54.6
ALFNet [12]	8.1	51.0
RepLoss [30]	5.0	47.9
CSP [9]	4.5	45.8
Proposed	**4.1**	**44.7**

Table 2. Comparison with the sate-of-the-art pedestrian detectors on Citypersons validation set in terms of MR^{-2} (%).

Method	Backbone	Stage	R	HO
FRCNN [31]	VGG16	2	15.4	-
FRCNN+Seg[31]	VGG16	2	14.8	-
TLL+MRF [32]	ResNet50	1	14.4	52.0
OR-CNN [33]	VGG16	2	12.8	55.7
RepLoss [30]	ResNet50	2	13.2	56.9
ALFNet [12]	ResNet50	1	12.0	51.9
CSP [9]	ResNet50	1	11.0	49.3
PRNet [10]	ResNet50	1	10.8	42.0
MGAN+ [27]	VGG16	2	11.0	39.7
KGSNet [28]	ResNet50	2	11.0	39.7
Ours	ResNet50	1	**10.6**	**36.7**

Table 3. Comparison of MR^{-2} (%) using the proposed feature extraction module with the baseline CSP on Caltech test set and Citypersons validation set.

Method	Feature Combination	Caltech		Citypersons	
		R	HO	R	HO
CSP	Concate-nation	6.8	50.7	11.7	41.8
Ours	Deformable trasformer encoder	4.1 (2.7↓)	44.7 (6.0↓)	10.6 (1.1↓)	36.7 (5.1↓)

5. CONCLUSIONS

A novel multi-scale deformable transformer encoder-based detection neck is proposed to improve the quality of the feature fed into the detection head in single-stage pedestrian detectors. The extracted feature is scale-aware with more attention on effective features and reduced attention on background information. With the proposed module, significant improvements in detecting reasonable and particularly heavily occluded pedestrians are observed on both Caltech and Citypersons datasets. This indicates the effectiveness and robustness of the proposed architecture. In the future, different uses of multi-scale features extracted by the encoder will also be studied.

6. REFERENCES

[1] N. Dalal and B. Triggs, "Histograms of Oriented Gradients for Human Detection," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 886-893 vol. 1, 2005.

[2] J. Yuan, P. Barmpoutis, and T. Stathaki, "Pedestrian Detection using Integrated Aggregate Channel Features and Multitask Cascaded Convolutional Neural-Network-based Face Detectors," *Sensors,* vol. 22, no. 9, pp. 3568, 2022.

[3] P. Dollar, C. Wojek, B. Schiele, and P. Perona, "Pedestrian Detection: An Evaluation of the State of the Art," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, no. 4, pp. 743-761, 2012.

[4] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A Diverse Dataset for Pedestrian Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 4457-4465, 2017.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 39, no. 6, pp. 1137-1149, 2017.

[6] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-Guided Attention Network for Occluded Pedestrian Detection," *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4967-4975, 2019.

[7] C. Chi, S. Zhang, J. Xing, Z. Lei, S. Z. Li, and X. Zou, "Pedhunter: Occlusion Robust Pedestrian Detector in Crowded Scenes," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, pp. 10639-10646, 2020.

[8] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable Transformers for End-to-End Object Detection," *arXiv preprint arXiv:.04159,* 2020.

[9] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detection," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5187-5196, 2019.

[10] X. Song, K. Zhao, W.-S. Chu, H. Zhang, and J. Guo, "Progressive Refinement Network for Occluded Pedestrian Detection," *European Conference on Computer Vision (ECCV)*, Springer, pp. 32-48, 2020.

[11] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:.02767,* 2018.

[12] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen, "Learning Efficient Single-Stage Pedestrian Detectors by Asymptotic Localization Fitting," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 618-634, 2018.

[13] Y. Tan, H. Yao, H. Li, X. Lu, and H. Xie, "PRF-Ped: Multi-scale Pedestrian Detector with Prior-based Receptive Field," *2020 25th International Conference on Pattern Recognition (ICPR)*, pp. 6059-6064, 2021.

[14] H. Xia, H. Wan, J. Ou, J. Ma, X. Lv, and C. Bai, "Mafa-Net: Pedestrian Detection Network Based on Multi-Scale Attention Feature Aggregation," *Applied Intelligence,* 2021.

[15] B. Ruan and C. Zhang, "Occluded Pedestrian Detection Combined with Semantic Features," *IET Image Process,* vol. 15, no. 10, pp. 2292-2300, 2021.

[16] Y. Xu and Q. Yu, "Adaptive Weighted Multi-Level Fusion of Multi-Scale Features: A New Approach to Pedestrian Detection," *Future Internet,* vol. 13, no. 2, p. 38, 2021.

[17] J. Ma, H. Wan, J. Wang, H. Xia, and C. Bai, "An Improved Scheme of Deep Dilated Feature Extraction on Pedestrian Detection," *Signal, Image and Video Processing,* vol. 15, no. 2, pp. 231-239, 2021.

[18] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," *arXiv preprint arXiv:2010.11929,* 2020.

[19] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-To-End Object Detection with Transformers," *European Conference on Computer Vision (ECCV)*, Springer, pp. 213-229, 2020.

[20] S. Paul and P.-Y. Chen, "Vision Transformers are Robust Learners," *arXiv preprint arXiv:2105.07581,* 2021.

[21] T.-Y. Lin *et al.*, "Microsoft Coco: Common Objects in Context," *European Conference On Computer Vision (ECCV)*, Springer, pp. 740-755, 2014.

[22] M. Lin *et al.*, "DETR for Crowd Pedestrian Detection," *arXiv preprint arXiv:.06785,* 2020.

[23] S. Shao *et al.*, "Crowdhuman: A Benchmark for Detecting Human in a Crowd," *arXiv preprint arXiv:1805.00123,* 2018.

[24] S. Zhang, R. Benenson, M. Omran, J. Hosang, and B. Schiele, "How Far Are We from Solving Pedestrian Detection?," *Proceedings of The IEEE Conference On Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, pp. 1259-1267, 2016.

[25] I. Hasan, S. Liao, J. Li, S. U. Akram, and L. Shao, "Generalizable Pedestrian Detection: The Elephant in the Room," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11328-11337, 2021.

[26] A. Tarvainen and H. Valpola, "Mean Teachers are Better Role Models: Weight-Averaged Consistency Targets Improve Semi-Supervised Deep Learning Results," *Advances in neural information processing systems,* vol. 30, 2017.

[27] J. Xie, Y. Pang, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao, "Mask-Guided Attention Network and Occlusion-Sensitive Hard Example Mining for Occluded Pedestrian Detection," *IEEE Transactions on Image Processing,* vol. 30, pp. 3872-3884, 2020.

[28] Y. Zhang, Y. Bai, M. Ding, S. Xu, and B. Ghanem, "KGSNet: Key-Point-Guided Super-Resolution Network for Pedestrian Detection in the Wild," *IEEE Transactions on Neural Networks and Learning Systems,* vol. 32, no. 5, pp. 2251-2265, 2021.

[29] L. Zhang, L. Lin, X. Liang, and K. He, "Is Faster R-Cnn Doing Well for Pedestrian Detection?," *European Conference on Computer Vision (ECCV)*, Springer, pp. 443-457, 2016.

[30] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion Loss: Detecting Pedestrians in a Crowd," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7774-7783, 2018.

[31] S. Zhang, R. Benenson, and B. Schiele, "CityPersons: A Diverse Dataset for Pedestrian Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[32] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu, "Small-Scale Pedestrian Detection Based on Topological Line Localization and Temporal Feature Aggregation," *European Conference on Computer Vision (ECCV)*, Springer International Publishing, Cham, pp. 554-569, 2018.

[33] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li, "Occlusion-Aware R-Cnn: Detecting Pedestrians in a Crowd," *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 637-653, 2018.