Deformable Vision Transformer Based Single-Stage Pedestrian Detector

Anonymous Author(s)
Affiliation
Address
email

Abstract

Pedestrian detection is a challenging field in computer vision, which requires both fast inference and high accuracy. Single-stage detectors are faster than region of interest based two-stage detectors, but the latter categorys decreased accuracy is mainly due to the lack of spatially adaptive features. To this end, we propose a single-stage anchor-free pedestrian detector with enhanced spatial and multi-scale features based on the deformable vision transformer aiming to achieve the balance between speed and accuracy. The architecture design is investigated in depth and comprehensive comparisons with state-of-the-art single- and two-stage detectors on various pedestrian datasets are performed. The proposed detector achieves leading performance on both Caltech and Citypersons datasets among single- and two-stage methods using less parameters compared to the baseline. The miss rates for R (3.8%) and HO (36.5%) are decreased to 2.6% and 28.0% on Caltech, and 10.6% and 36.7% on Citypersons validation datasets respectively. It even outperforms SOTA two-stage detectors in HO subset by 3% on Citypersons validation set.

1 Introduction

Pedestrian detection is a popular task subordinate to object detection in computer vision. This task aims at accurately locating and classifying pedestrians in images or videos obtained via visible-light and infrared cameras. In this paper, we focus on detection using RGB images. Pedestrian detection is a significant procedure as it is serving as the prerequisite for various vision tasks [1], such as human-centric tasks (person re-identification [2, 3], person search [4], human pose estimation [5]) and more generic multi-object tracking [6]. In this case, it plays an important role in autonomous driving [7, 8], video surveillance [9], and action tracking.

Pedestrian detection is a challenging task, mainly due to the varying appearances resulting from the significant occlusions and the varying scales. On the one hand, significant intra- or inter- class occlusion occurs when a pedestrian is occluded by other pedestrians or objects like cars and bicycles in busy areas. Due to heavy occlusions, only partial human body is visible which significantly reduces the discriminative features. On the other hand, when the target is far away from the camera, its height and width decrease in the image, and vice versa. Large pedestrians tend to have more informative features but are difficult to be fully extracted from a vast region. Small targets are compact but relatively ambiguous with less preserved features. In summary, the fluctuating amount and varying shape of effective features are the key problems. These intrinsic problems challenge the capabilities of the feature extraction modules, which act as a long-standing bottleneck in pedestrian detection and almost all other object detection tasks.

To deal with these problems, deep Convolutional Neural Networks (CNNs)-based two-stage methods [10, 11, 12, 13, 14, 15] have pervaded pedestrian detection tasks owing to the high detection

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

accuracy. These methods first make a coarse prediction of targets in the Region Proposal Network (RPN), features inside which are then utilized to refine the bounding boxes and predict the final score. However, the repeated prediction requires longer training/inference time. Secondly, the design of the anchors in RPN depends substantially on the choice of parameters. Lastly, jointly training two networks is complicated.

To achieve faster inference and easier training procedures, single-stage approaches, which only make a single round prediction, are developed [16] and have been applied in pedestrian detection [17, 18, 19, 20, 21] prediction. Single-stage methods are promising; however, they suffer from decreased detection accuracy. The miss rates of two-stage methods in reasonable subset on Caltech are smaller than 4% [14][10, 22, 23], while those for single-stage methods are larger than 4.5% [19]. For Citypersons validation set, the miss rates for the former is less than 40% in HO subset which is much lower than the latter (best result 42% [21]). Our research purpose is to improve the detection performance of the single-stage pedestrian detectors, especially in R and HO subsets, to narrower the gap between single- and two-stage methods.

Massive studies have been carried out aiming to enhance the single-stage pedestrian detectors based on Center and Scale Prediction (CSP) [19], a typical single-stage and anchor-free pedestrian detector which significantly accelerates the inference [1]. In the past year, most research were focusing on the fusion of representative features at multiple levels [24, 25, 26, 27, 28, 29]. [27] enhanced features at different levels via increasing semantic information at the low-level and enriches the localization information at the high-level. Similarly, [29] fused multi-level feature maps in adequate proportions. [26, 28] also explored multi-level feature aggregation. These enhancements make sense for the intrinsically unbalanced feature information between shallower and deeper feature maps. However, this unbalance problem also exists in two-stage detectors. This means that it is not the root cause to the poorer performance of single-stage methods in comparison to two-stage methods.

To figure out the root cause, we compared the general architecture of single- and two-stage detectors in Figure 1. Assume that the training strategies and the detection head are the same for both methods, the difference in performance lies in the information fed into the detection head. For two-stage methods, both the proposal region positions and features within this domain are fed into the detection head. Thereby, the detection head classifies via spatially target-focused features with less background interference and refines the bonding boxes based on region positions. For single-stage methods, each pixel in the feature map serves as the proposal regionwithout pre-estimated positions. Features with these proposal regions share the same receptive field size. This may be too small to include sufficient information of large targets, or the receptive field is so large that the background information submerges the visible pedestrians. Obviously, this is more challenging for classifiers compared to the two-stage method. As for the proposal region positions, single-stage methods have to regress from scratch, which is also more difficult than simple refinement in two-stage methods. These suggest that the lack of spatially target-focused feature representation and the position prediction from scratch as two key bottlenecks hindering the progress of single-stage detectors.

Thus, to focus features on the targets or other useful information automatically without the assistance of region proposals, we take the advantage of vision transformers. Vision transformers use attention weights to describe the pairwise dependency of each entity in the feature map. They output the weighted averaged feature which aggregates important entities while supressing the disturbing information. With attention mechanism on top of the backbone spatially adaptive features are supplemented which benefits the subsequent classification and regression tasks. In this case, the enhanced detector can make the best use of both the fast inference originated from single-stage methods and more effective and cleaner features. Our main contributions are as follows:

- Propose a novel multi-scale deformable feature extractor based on a deformable vision transformer. The features fed into the detection head is enhanced along both scale and spatial dimensions.
- Propose a new single-stage detector based on the proposed feature extractor and CSP. Comprehensive experiments on multiple pedestrian datasets show that the performance gap between the presented method and two-stage methods is narrowed.
- Extend the application of vision transformers on top of the backbone in pedestrian detection tasks.

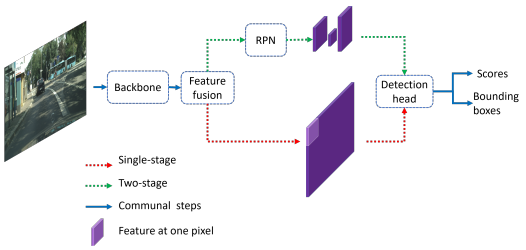

Figure 1: Generic detection pipelines of the single- and two-stage detectors. The detection head consists of two parallel branches to classify and regress.

2 Related works

The vision transformer in Computer Vision (CV) originates from the transformers widely utilized in Natural Language Processing (NLP). Thanks to the attention mechanism, vision transformers are applied to image feature extraction as the backbone (like ViT [30] and Swin Transformer [31]) or on top of the backbone (like DETR [32]). Our detector belongs to the latter category. DETR consists of the convolutional backbone, six encoder-decoder pairs and the prediction head. It is an inspiring end-to-end detector which has aroused plenty of studies based on it. One typical variant is deformable DETR [33] which outperforms Faster R-CNN [34] and DETR on COCO 2017 validation set [35]. It is known that DETR requires massive memory to store the self-attention weights within each Multi-Head Self-Attention (MHSA) layer. The memory cost is linear to the number of attention heads and is square to the number of pixels in the feature map. Additionally, first- and second-order momentum in optimizers introduce in further memory cost in training. For accurate pedestrian detection, more attention heads and the relatively large feature map are preferred, which results in memory explosion. To this end, the fast-converging and memory-saving deformable DETR is proposed. It calculates the attention weights at several sampling locations rather than each pixel in the feature map. Its memory cost is linear to the number of pixels, which makes it possible to train with high resolutions but limited memory.

So far, vision transformer seems inspiring and potential, but it is rarely applied in pedestrian detection in the form of DETR or its variants. [36] found that deformable DETR not only behaves worse than Faster R-CNN on CrowdHuman dataset [37], but also costs tenfold training time. Although dense object queries and the rectified attention field are proposed to narrow the performance gap, the modified deformable DETR still shows a limited advantage over the traditional Fast R-CNN. Different from the use of six encoder-decoder pairs in deformable vision transformer,BoTNet [38] achieves competing performacne on ImageNet validation set [39] with much simpler usage of attention machanism. It substitutes the convolutional layers in the last stage of ResNet [40] for MHSA layers. This implies that rigidly utilizing the whole six encoder-decoder pairs in the pedestrian detector may not be cost-effective. Inspired by this, our work utilizes one deformable encoder to enhance the single-stage pedestrian detector.

3 Method

Deformable transformer encoder: Figure 2 shows the deformable encoder in the multi-scale deformable vision transformer. This encoder takes the multi-scale feature maps $\left\{\boldsymbol{z}^l\right\}_{l=1}^{L}$ ($L = 3$ is the number of feature maps) with height H_l and width W_l at scale l and reference points, which are the positions of grid centers of the maps, as inputs. It outputs the enhanced feature maps with the same resolution as the input. The input feature maps are initially added by fixed encoded positional [32] and learnable level information [33] to disambiguate spatial and scale positions. Then they are projected to the query feature map $\boldsymbol{z}_q$ via linear projection. The input feature maps generate value feature maps $\boldsymbol{z}_v$ with a linear layer but no positional or level encodings. The query feature map $\boldsymbol{z}_q$, value feature map $\boldsymbol{z}_v$ together with the pre-generated reference points are sent to the Multi-Scale Deformable Attention (MSDA) layer to produce enhanced spatially adaptive features. The MSDA layer is followed by a Fee-Forward Network (FFN) which is a fixed module in common transformer encoders. In summary, the encoder supplements the semantic information of input feature maps via the embedded attention layer and FFN.

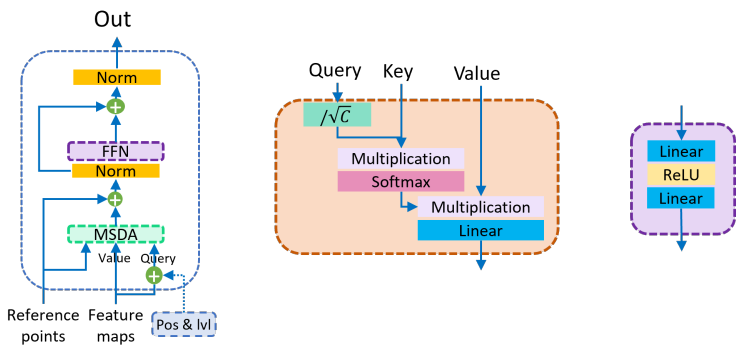

Figure 2: Structures of the deformable encoder in the multi-scale deformable vision transformer and embedded multi-scale deformable attention (MSDA) layer. **Left**: The deformable encoder is the same as the encoder in DETR except for that MSDA does not need key tensor while reference points are required instead. **Mid**: The MSDA averages the features at sparse sampling locations on different scales with weights computed from the query tensor. **Right**: Feed-Forward Network (FFN).

Multi-scale deformable attention: The MSDA layer sums the entities selected from the value feature map $\boldsymbol{z_v}$. Selected feature entities locate at predicted sampling locations with deformable attention weights. The deformable attention weights $\boldsymbol{W}$ in MSDA are the linear projection of query features $\boldsymbol{z_q}$ followed by a softmax operator along the scale and sampling point dimensions. For a single query entity, it only needs $N_h N_l N_p$ attention weights, representing the significance of selected value features at different attention heads, scales, and points. Sampling locations are the summation of reference points $\boldsymbol{p}$ and sampling offsets $\Delta\boldsymbol{p}$ which is the linear embedding of the query features. For float sampling locations, selected value feature is bilinear interpolated for accuracy and training offset predictor. The q-th feature vector of the separate output feature $\boldsymbol{z}^{\boldsymbol{so},h} \in \mathbb{R}^{N_q \times c_v} (N_q = \sum_{l=1}^{L} H_l W_l)$ at attention head h (total N_h heads) is

$$\boldsymbol{z}_q^{\boldsymbol{so},h} = \sum_{p=1}^{N_p} \sum_{l=1}^{L} W_{plhq} \boldsymbol{v}_{\boldsymbol{p}_{ql} + \Delta \boldsymbol{p}_{plhq}} \tag{1}$$

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

$$\boldsymbol{z}_q^{\boldsymbol{o}} = \sum_{h=1}^{N_h} \boldsymbol{W}'_h \boldsymbol{z}_q^{\boldsymbol{so},h} \tag{2}$$

where $\boldsymbol{W}'_h \in \mathbb{R}^{c \times c_v}$ denotes the learnable weight for the h-th attention head.

Center and scale prediction: The architecture of CSP is simple: It concatenates multi-scale feature maps and fed them to the detection head. Two main innovations of CSP are the anchor-free structure and predicting the center heatmap, height map and the optional offsets w.r.t. the centers instead of the traditional bounding boxes. Note that "scale map" is used in the original work, however, we use "height map" in case of confusion with the feature map scale. With no anchor applied, less hyper-parameters are required which simplifies training and design. The sample center and scale ground truths are shown in Figure 3. Centers of all pedestrians are assigned as positives in the center heatmap, while negatives for other positions. For the height map, these points are assigned with the logarithm of the pedestrian height.

Proposed single-stage detector with enhanced features: The proposed detector consists of the ResNet 50 backbone, the detection neck, and the head as shown in Figure 3. The detection head predicts the center and scale (height) maps and offsets to slightly adjust the position of pedestrians. The

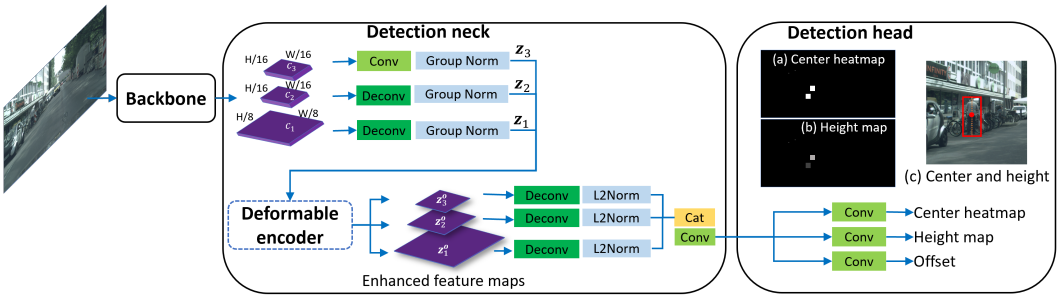

Figure 3: The overall architecture of the proposed advanced single-stage pedestrian detector. The input images are fed into the backbone, followed by the detection neck with a deformable encoder based feature extractor to enhance both the spatial and multi-scale information. The detection head is the same as CSP. **(a)**: The ground truth center heatmap highlights pedestrian centers with white squares. **(b)**: The ground truth scale map shows the logarithm of the height. **(c)**: Sketch of the center (red node) and height (orange dotted line) ground truths and their relation to the bounding box (red box).

features are enhanced in the detection neck where we take the advantage of the deformable transformer encoder, particularly the embedded MSDA mechanism to supplement the spatially adaptive features across different scales. In the neck, input feature maps are first upsampled with deconvolutional layers or encoded with the convolutional layer to generated multi-scale feature maps $\{\boldsymbol{z}^l\}_{l=1}^{3}$. They are followed by group normalization to prevent the Internal Covariate Shift (ICS) that might be inducted by subsequent linear operations in the deformable encoder. The encoder yields enhanced multi-scale feature maps. Enhanced features are upsampled to keep the resolution as $(H/4, W/4)$ for accurate detection. They are normalized via L2Norm [19] before concatenation along the channel dimension. This force the features at different scales to contribute equally to the feature representation. Concatenated features are compressed along the channel dimension to reduce the network parameters. Compared to the architecture of CSP, the proposed detector adopts the novel multi-scale deformable attention mechanism in the detection neck. Furthermore, an additional convolutional layer is applied in CSP detection head ahead of the branches.

Training: Same ground truth and loss functions as [19] are utilized. The overall loss consists of three parts:

$$L = \lambda_c L_{center} + \lambda_s L_{height} + \lambda_o L_{offset} \tag{3}$$

where the center heatmap loss L_{center}, height map loss L_{height} and offset loss L_{offset} are given in the supplementary. Weights for each loss λ_c, λ_s and λ_o are set as 0.01, 1 and 0.1.

Inference: Predicted scores above 0.01 are kept. The predicted pedestrian centers are the summation of the offsets and positions of those qualified scores. The predicted width is the height multiplied by the constant 0.41 [19]. The predicted sizes are remapped to the original image size and fed to NMS.

4 Experiments

4.1 Settings

Datasets: The proposed detector is evaluated on two commonly used public pedestrian datasets: **Caltech** [41] and **Citypersons** [42]. The Caltech dataset is an approximately 10 hours of 480x640 video taken in a single urban city. The standard training set contains about 250,000 frames with 350,000 bounding boxes. In our experiments, the training data augmented by 10 folds containing 42782 images with 13674 persons and the standard test set containing 4024 images with corresponding new annotations [43] and fixed aspect-ratio for bounding boxes [44] are used. Citypersons training set was recorded across 18 different cities, 3 seasons and various weather conditions with 19654 persons in 2975 high resolution images (1024x2048). The validation set contains 500 images across 3 cities.

Training details: Our network is established on [44]. ResNet 50 pretrained on ImageNet with Adam, moving average weights [45] and step learning rate schedule are applied. Data augmentation

Table 1: The experimental settings of four subsets.

Subset	Reasonable	Reasonable Small	Heavily Occluded	All
Height	[50, inf]	[50, 70]	[50, inf]	[20, inf]
Visibility	[0.65, inf]	[0.65, inf]	[0.2, 0.65]	[0.2, inf]

Table 2: The experimental settings of four subsets.

z^1	z^2	z^3	R	HO
1/4	1/8	1/16	**4.1**	**44.7**
1/4	1/16	1/16	4.2	45.9
1/8	1/8	1/16	4.6	48.5
1/4	1/8	1/8	5.9	47.8

techniques including random horizontal flip and scaling are applied. For Caltech, additional random color distortion and cropping are implemented. The input images are rescaled to 336x448 for Caltech and 640x1280 for Citypersons dataset. The proposed network is trained with a single NVIDIA GeForce RTX 3090 GPU for 10 and 75 epochs with batch size 16 and 4 on Caltech and Citypersons respectively. The learning rate and related hyperparameters will be presented afterwards according to different designs of architectures. All the training is performed with a randomly chosen and fixed seed.

Metrics: Log-average Miss Rate (denoted as MR^{-2}) over False Positive Per Image (FPPI) ranging in $[10^{-2}, 10^{0}]$ is calculated over Reasonable (R), Reasonable Small (RS), Heavily Occluded (HO), and All (A) subsets defined in Table 1. The lower the MR^{-2} the better.

4.2 Ablation studies

In this section, ablation studies are conducted to investigate the influence of network architectures on the detection performance. All studies are performed on the Caltech dataset.

Scales of feature maps: Different combinations of multi-scale feature maps $\{\boldsymbol{z}^l\}_{l=1}^{3}$ with three downsample ratios ($1/4$, $1/8$ and $1/16$) are compared in Table 2. In this comparison, only $\boldsymbol{z}_3^o$ which has the same resolution as $\boldsymbol{z}^3$ upsampled to $(H/4, W/4)$ and fed into the detection head without intermediate concatenation and L2Norm. Results show that feature maps with ratios $1/4$, $1/8$, and $1/6$ for each scale produce the best performance. In this case, we choose to upsample $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ by two times while maintain the scale of $\boldsymbol{c}_3$ in the proposed detector and the following experiments.

Scales of enhanced feature maps: Different combinations of multi-scale enhanced feature maps $\{\boldsymbol{z}_l^o\}_{l=1}^{3}$ are fed into the detection head. Note that the downsample ratios for them are $1/4$, $1/8$ and $1/16$ respectively. All the enhanced feature maps are first upsampled to $(H/4, W/4)$, if needed, followed by normalization if multiple feature maps are utilized. They are then processed in three ways: 1. Cat: Concatenate along the channel dimension followed by a compression layer to reduce the number of channels to 256. 2. Add: Add them elementwise. 3. Sep: No fusion across multi-scales but send them separately to the detection head which doubles the number of predictions. Table 3 compared various strategies to fuse the multi-scale enhanced feature maps. [1] train boldly without any learning rate schedules. It shows that concatenation followed by L2Norm produces the best results in both R and HO subsets on Caltech test set.

Choice of normalization method applied to $\{\boldsymbol{c}^l\}_{l=1}^{3}$: As Table 3 shows, GN performs best in R subset while L2N in HO subset. Considering that the miss rate of GN is 11.6% lower in R subset and only 1.2% higher in HO subset compared to L2N, GN is utilized in our experiments as indicated in Figure 3.

Number of encoders: Different numbers of encoders are applied in the proposed architecture. These encoders are connected in series. Table 4 presents that although more encoders can provide higher-level semantic information, the best result is observed when a single encoder is applied. This phenomenon supports the design of BotNet [38] to some extent, which indicates that following the whole set of six encoder-decoder pairs in (deformable) DETR may not be suitable for specific

Table 3: Combinations of enhanced feature maps with different scales in different ways followed by three Normalization Methods (NM). L2N, LN and GN stand for L2Norm, Layer Normalization and Group Normalization respectively.

z_1^o	✓			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
z_2^o		✓												✓
z_3^o			✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Cat							✓	✓	✓	✓	✓	✓	✓	✓
Add				✓	✓									
Sep						✓								
NM[2]					L2N	L2N		GN[1]	LN[1]	L2N[1]	L2N	L2N	L2N	L2N
NM[3]	GN	GN	GN	GN	GN	GN	GN	GN	GN	GN	GN	L2N	BN	GN
R	4.3	5.7	4.1	7.0	**3.7**	4.2	5.9	6.4	4.0	3.9	**3.7**	4.3	4.5	5.3
HO	44.5	48.8	44.7	51.3	45.5	48.1	50.7	49	47.4	44.5	42.8	**42.6**	44.6	44.1

Table 4: Using 1 to 4 encoders in the detection neck.

Number of encoders	1	2	3	4
R	**3.7**	4.9	4.8	5.5
HO	**42.8**	46.6	43.1	45.1

object detection tasks. A single encoder can work effectively in pedestrian detection task with the least number of learnable parameters which prevents overfitting.

Significance of enhanced feature maps: We evaluated the baseline CSP detector and the proposed detector which improve CSP by adopting the deformable encoder-based feature enhancement module in the detection neck. Both the single enhanced feature map and the concatenation of two maps are assessed. Evaluation is conducted on Caltech (Table 5) and Citypersons (Table 6) respectively. Results from both datasets indicate that with the enhanced feature map(s), the miss rates are decreased significantly in R, HO and A subsets, in particular, the HO subset witnesses a decrease up to 7.9%, and R subset up to 3.1%. This means that the idea of extracting spatially adaptive features works effectively for the single-stage detector. With the concatenation of multi-scale enhanced feature maps (row three), the performance is further improved comprehensively on Caltech dataset compared to the single feature map (row two). Due to the additional interpolation operation in deformable self-attention layer, the proposed detectors require longer inference time compared to the baseline, however, the proposed detector contains less learnable parameters as shown in Table 7.

4.3 Comparison with the state-of-the-art method

We compared two proposed architectures, namely the single enhanced feature map z_3^o and the concatenation of z_3^o and z_1^o, with the state-of-the-art single- and two-stage detectors on Caltech and Citypersons datasets. Proposed detectors are trained with the hyperparameters presented in Table 8. As Table 9 and Table 10 show, the proposed detectors achieve lowest miss rates in R and HO subsets on both datasets among listed single-stage detectors, and leading performances in A subset. For Caltech dataset, the lowest miss rate (3.7%) of the proposed detector in R subset is 0.2% smaller than that of the two-stage KGSNet as presented in the upper part of Table 9. With pretraining on Citypersons dataset, the miss rates on all three subsets of Caltech are considerably reduced and reach the lowest compared to other pretrained detectors shown in the bottom part of Table 9. For Cityper-

Table 5: Comparison of the baseline CSP detector and the proposed detector with enhanced features and additional concatenated feature maps on Caltech test set. FPS stands for frames per second evaluated during inference process with input image size 480x640.

	R	HO	A	FPS
Baseline	6.8	50.7	62.3	39.2
+Enhanced features	4.1 (2.7↓)	44.7 (6.0↓)	57.9 (4.4↓)	32.1
++Cat z_1^o,z_3^o	3.7 (3.1↓)	42.8 (7.9↓)	56.97 (5.3↓)	29.5

Table 6: Comparison of the baseline CSP detector and the proposed detector with enhanced features and additional concatenated feature maps on Citypersons validation set. FPS stands for frames per second. FPS stands for frames per second evaluated during test process with input image size 1024x2048.

	R	RS	HO	A	FPS
Baseline	11.7	14.4	41.8	38.2	8.2
+Enhanced features	10.6 (1.1↓)	14.9 (0.5↑)	36.7 (5.1↓)	36.4 (1.8↓)	7.3
++Cat z_1^o,z_3^o	10.9 (0.8↓)	13.7 (0.7↓)	38.6 (3.2↓)	37.2 (1.0↓)	6.8

Table 7: The number of learnable parameters in the baseline and proposed detectors.

	Backbone	Neck	Head	Total Param
Baseline	23.51M	14.68M	1.77M	39.96M
+Enhanced feature		8.60M	1.03K	32.11M
++Cat z_1^o,z_3^o		8.73M		32.24M

Table 8: Hyperparameters for training the proposed detectors on Caltech and Citypersons datasets respectively. The training starts from the base learning rate l_r which drops to $r\times$ after s epochs.

Dataset	Type	l_r	s	r
Caltech	Single	1e-4	5	0.2
	Cat	0.5e-4	6	0.5
Citypersons	Single	2e-4	40	0.4
	Cat	0.5e-4	60	0.5

Table 9: Comparison with the state-of-the-art single- and two-stage pedestrian detectors on Caltech test set. † Results are cited from [19] for completeness. * Concatenated enhancned feature maps. Bottom part:Pretrained on Citypersons and tested on Caltech. The best and second-best results are in bold and underlined. Red for two-stage methods and black for single-stage methods in the upper part of the table.

Method	Backbone	Stage	R	HO	A
Faster R-CNN† [34]	ResNet50	2	8.7	53.1	62.6
ALFNet† [18]	ResNet50	1	8.1	51.0	59.1
RPN+BF† [46]	VGG16	2	7.3	54.6	59.9
RepLoss† [14]	ResNet50	2	5.0	47.9	59.0
CSP [19]	ResNet50	1	4.5	45.8	**56.9**
KGSNet [23]	ResNet50	2	3.9	**34.2**	**42.2**
JointDet [47]	ResNet50	2	3.0	-	-
PedHunter [10]	ResNet50	2	**2.3**	-	-
Proposed	ResNet50	1	4.1	44.7	57.9
Proposed*	ResNet50	1	**3.7**	**42.8**	56.97
ALFNet† [18]	ResNet50	1	4.5	43.4	56.8
RepLoss† [14]	ResNet50	2	4.0	41.8	58.6
CSP [19]	ResNet50	1	3.8	36.5	54.4
Proposed	ResNet50	1	3.2	32.2	56.5
Proposed*	ResNet50	1	**2.6**	**28.0**	**53.9**

Table 10: Comparison with the state-of-the-art pedestrian detectors on Citypersons validation set. * Concatenated enhancned feature maps. Results are cited from [19] for completeness.[1] Less than 65% visibility instead of 20-65% visibility. Best and second-best results are in bold and underlined.

Method	Backbone	Stage	R	HO	A
FRCNN† [42]	VGG16	2	15.4	-	-
FRCNN+Seg†[42]	VGG16	2	14.8	-	-
TLL+MRF [20]	ResNet50	1	14.4	52.0	-
OR-CNN [15]	VGG16	2	12.8	55.7[1]	-
RepLoss [14]	ResNet50	2	13.2	56.9[1]	-
ALFNet [18]	ResNet50	1	12.0	51.9	-
CSP [19]	ResNet50	1	11.0	49.3	-
PRNet [21]	ResNet50	1	<u>10.8</u>	42.0	-
MGAN+ [48]	VGG16	2	**11.0**	**39.7**	-
KGSNet [23]	ResNet50	2	**11.0**	**39.7**	**36.2**
Proposed	ResNet50	1	**10.6**	**36.7**	<u>36.4</u>
Proposed*	ResNet50	1	10.9	<u>38.6</u>	37.2

sons dataset, the proposed two detectors even outperform the competitive two-stage detectors in R, and especially HO subset with miss rate 36.7% which is 3% lower than that of the best of two-stage detectors. Overall, it can be observed that with the enhanced spatial adaptive and multi-scale features, the gap between single- and two-stage detectors in R and HO subsets on different datasets has been narrowed. Surprisingly, the proposed single-stage method can produce better results than accurate two-stage methods on certain pedestrian dataset, such as Citypersons dataset. In this case, our enhanced single-stage detector not only predicts quickly, but also with accuracy comparable to complicated two-stage methods.

5 Conclusion

We propose a single-stage anchor-free pedestrian detector with enhanced spatial and multi-scale features. This is fulfilled by the combination of deformable self-attention mechanism of the deformable vision transformer and Center and Scale Prediction (CSP). Comprehensive experiments over architecture design in the detection neck and comparison with the state-of-the-arts are performed. More than 40% parameters are reduced in the detection neck compared to CSP; the detector achieves the best results in R and HO subsets among all the listed single-stage detectors on both Caltech and Citypersons datasets. With pretraining, the miss rates for these subsets can be decreased to 2.6% and 28.0% which is far better than other single-stage methods and comparable to two-stage methods (2.3% for R). It even outperforms SOTA two-stage detectors in HO subset by 3% miss rate on Citypersons validation set. The proposed detector significantly improves the accuracy of the single-stage method with slightly decreased inference speed. These results demonstrate that single-stage detectors can do better if spatially adaptive and multi-scale features are jointly adopted, thus making them cost-effective and efficient.

References

[1] Jiale Cao, Yanwei Pang, Jin Xie, Fahad Shahbaz Khan, and Ling Shao. From handcrafted to deep features for pedestrian detection: A survey. *IEEE transactions on pattern analysis and machine intelligence*, PP, 2021.

[2] Wei Li, Xiatian Zhu, and Shaogang Gong. Harmonious attention network for person re-identification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2285–2294.

[3] Yulin Li, Jianfeng He, Tianzhu Zhang, Xiang Liu, Yongdong Zhang, and Feng Wu. Diverse part discovery: Occluded person re-identification with part-aware transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2898–2907.

[4] Yichao Yan, Jinpeng Li, Jie Qin, Song Bai, Shengcai Liao, Li Liu, Fan Zhu, and Ling Shao. Anchor-free person search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699.

[5] Yanjie Li, Shoukui Zhang, Zhicheng Wang, Sen Yang, Wankou Yang, Shu-Tao Xia, and Erjin Zhou. Tokenpose: Learning keypoint tokens for human pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11313–11322.

[6] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Bridging the gap between detection and tracking: A unified approach. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3999–4009.

[7] Zhilu Chen and Xinming Huang. Pedestrian detection for autonomous vehicle using multi-spectral cameras. *IEEE Transactions on Intelligent Vehicles*, 4(2):211–219, 2019.

[8] Antonin Vobecky, Michal Uricár, David Hurych, and Radoslav Skoviera. Advanced pedestrian dataset augmentation for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

[9] Quanzheng Mou, Longsheng Wei, Conghao Wang, Dapeng Luo, Songze He, Jing Zhang, Huimin Xu, Chen Luo, and Changxin Gao. Unsupervised domain-adaptive scene-specific pedestrian detection for static video surveillance. *Pattern Recognition*, 118:108038, 2021.

[10] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Pedhunter: Occlusion robust pedestrian detector in crowded scenes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10639–10646.

[11] Chunze Lin, Jiwen Lu, Jie for Video Technology. Multi-grained deep feature learning for robust pedestrian detection. 29(12):3608–3621, 2018.

[12] T. Liu, W. Luo, L. Ma, J. J. Huang, T. Stathaki, and T. Dai. Coupled network for robust pedestrian detection with gated multi-layer feature extraction and deformable occlusion handling. *IEEE Transactions on Image Processing*, 30:754–766, 2021.

[13] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4967–4975.

[14] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen. Repulsion loss: Detecting pedestrians in a crowd. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7774–7783.

[15] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 637–653.

[16] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[17] Chunze Lin, Jiwen Lu, Gang Wang, and Jie Zhou. Graininess-aware deep feature learning for pedestrian detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 732–747.

[18] Wei Liu, Shengcai Liao, Weidong Hu, Xuezhi Liang, and Xiao Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 618–634.

[19] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. High-level semantic feature detection: A new perspective for pedestrian detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5187–5196.

[20] Tao Song, Leiyu Sun, Di Xie, Haiming Sun, and Shiliang Pu. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *European Conference on Computer Vision (ECCV)*, pages 554–569. Springer International Publishing.

[21] Xiaolin Song, Kaili Zhao, Wen-Sheng Chu, Honggang Zhang, and Jun Guo. Progressive refinement network for occluded pedestrian detection. In *European Conference on Computer Vision (ECCV)*, pages 32–48. Springer.

[22] Hong-hui Xu, Xin-qing Wang, Dong Wang, Bao-guo Duan, and Ting Rui. Object detection in crowded scenes via joint prediction. *Defence Technology*, 2021.

[23] 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*, 32(5):2251–2265, 2021.

[24] Jun Ma, Honglin Wan, Junxia Wang, Hao Xia, and Chengjie Bai. An improved one-stage pedestrian detection method based on multi-scale attention feature extraction. *Journal of Real-Time Image Processing*, 18(6):1965–1978, 2021.

[25] Jun Ma, Honglin Wan, Junxia Wang, Hao Xia, and Chengjie Bai. An improved scheme of deep dilated feature extraction on pedestrian detection. *Signal, Image and Video Processing*, 15(2):231–239, 2021.

[26] Binjie Ruan and Chongyang Zhang. Occluded pedestrian detection combined with semantic features. *IET Image Process*, 15(10):2292–2300, 2021.

[27] Y. Tan, H. Yao, H. Li, X. Lu, and H. Xie. Prf-ped: Multi-scale pedestrian detector with prior-based receptive field. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 6059–6064.

[28] Hao Xia, Honglin Wan, Jiayu Ou, Jun Ma, Xinyao Lv, and Chengjie Bai. Mafa-net: Pedestrian detection network based on multi-scale attention feature aggregation. *Applied Intelligence*, 2021.

[29] Yao Xu and Qin Yu. Adaptive weighted multi-level fusion of multi-scale features: A new approach to pedestrian detection. *Future Internet*, 13(2):38, 2021.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, and Sylvain Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022.

[32] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229. Springer.

[33] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:.04159*, 2020.

[34] 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*, 39(6):1137–1149, 2017.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference On Computer Vision (ECCV)*, pages 740–755. Springer.

[36] Matthieu Lin, Chuming Li, Xingyuan Bu, Ming Sun, Chen Lin, Junjie Yan, Wanli Ouyang, and Zhidong Deng. Detr for crowd pedestrian detection. *arXiv preprint arXiv:.06785*, 2020.

[37] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

[38] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16519–16529.

[39] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, and Michael Bernstein. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[40] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[41] Piotr Dollar, Christian Wojek, Bernt Schiele, and Pietro Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE transactions on pattern analysis and machine intelligence*, 34(4):743–761, 2011.

[42] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition*, pages 4457–4465.

[43] Shanshan Zhang, Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele. How far are we from solving pedestrian detection? In *Proceedings of The IEEE Conference On Computer Vision and Pattern Recognition*, pages 1259–1267.

[44] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Generalizable pedestrian detection: The elephant in the room. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11328–11337.

[45] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. *Advances in Neural Information Processing Systems*, 30, 2017.

[46] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He. Is faster r-cnn doing well for pedestrian detection? In *European Conference on Computer Vision (ECCV)*, pages 443–457. Springer.

[47] Cheng Chi, Shifeng Zhang, Junliang Xing, Zhen Lei, Stan Z Li, and Xudong Zou. Relational learning for joint head and human detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10647–10654.

[48] Jin Xie, Yanwei Pang, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network and occlusion-sensitive hard example mining for occluded pedestrian detection. *IEEE Transactions on Image Processing*, 30:3872–3884, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [No] The code will be published online in the future.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] The data are available online.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]