# Localized Semantic Feature Mixers for Efficient Pedestrian Detection in Autonomous Driving

Abdul Hannan Khan[1,2], Mohammed Shariq Nawaz[1], Andreas Dengel[1,2]
Department of Computer Science, RPTU Kaiserslautern-Landau[1],
German Research Center for Artificial Intelligence (DFKI GmbH)[2],
67663 Kaiserslautern, Germany
Corresponding Author: hannan.khan@dfki.de

## Abstract

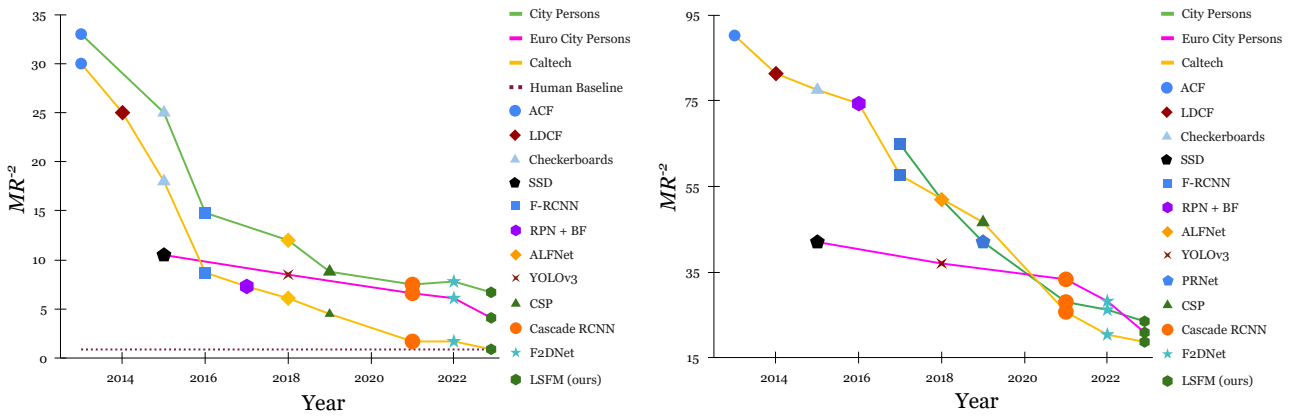
*Autonomous driving systems rely heavily on the underlying perception module which needs to be both performant and efficient to allow precise decisions in real-time. Avoiding collisions with pedestrians is of topmost priority in any autonomous driving system. Therefore, pedestrian detection is one of the core parts of such systems' perception modules. Current state-of-the-art pedestrian detectors have two major issues. Firstly, they have long inference times which affect the efficiency of the whole perception module, and secondly, their performance in the case of small and heavily occluded pedestrians is poor. We propose Localized Semantic Feature Mixers (LSFM), a novel, anchor-free pedestrian detection architecture. It uses our novel Super Pixel Pyramid Pooling module instead of the, computationally costly, Feature Pyramid Networks for feature encoding. Moreover, our MLPMixer-based Dense Focal Detection Network is used as a light detection head, reducing computational effort and inference time compared to existing approaches. To boost the performance of the proposed architecture, we adapt and use mixup augmentation which improves the performance, especially in small and heavily occluded cases. We benchmark LSFM against the state-of-the-art on well-established traffic scene pedestrian datasets. The proposed LSFM achieves state-of-the-art performance in Caltech, City Persons, Euro City Persons, and TJU-Traffic-Pedestrian datasets while reducing the inference time on average by 55%. Further, LSFM beats the human baseline for the first time in the history of pedestrian detection. Finally, we conducted a cross-dataset evaluation which proved that our proposed LSFM generalizes well to unseen data.*

## 1. Introduction

Autonomous driving is currently under the spotlight in the computer vision community [3, 20]. Detecting and avoiding collisions with pedestrians is one of the numerous challenges of autonomous driving. Pedestrian detectors for autonomous driving not only have to be performant but efficient as well, since rapid perception is required to make timely decisions. Furthermore, these systems need to fulfill additional constraints such as good portability and low computational footprint, as compute-intensive systems can have a heavy impact on the milage of autonomous vehicles.

Pedestrian detection for autonomous driving aims to provide the autonomous vehicle with a timely perception of all pedestrians in its surroundings. The problem becomes more challenging as most of the pedestrians are occluded either by other pedestrians or by other objects [5,48]. Additionally, the camera stream is introduced with motion blur since it is coming from the camera mounted on a moving vehicle [15]. The motion blur problem further intensifies when the vehicle moves faster. Dealing with motion blur and occlusion is vital for a pedestrian detector to perform well. Another major challenge for pedestrian detectors is the scale variance in pedestrians. Since the camera images are subject to perspective distortion the pedestrian scales vary from a few pixels large to almost equal to the height of the image frame. Small-scale pedestrians (far or short) are the bottleneck of scale variance problem [5]. The pedestrian detector needs to sufficiently understand the core visual features of a pedestrian and use them to detect pedestrians irrespective of their scales.

Furthermore, domain generalization is critical for a pedestrian detector as it is expected to perform in all circumstances *e.g.*, all kinds of weather, lighting, and traffic densities, which might or might not be part of the training data [14, 15]. Therefore, pedestrian detectors should perform well on unseen data to be reliable under real-world circumstances.




(a) Reasonable (b) Heavy Occlusion

Figure 1. Performance of pedestrian detectors in different settings and their evolution over the years. Both figures contain data on three different pedestrian detection datasets namely, City Persons [48] (Green), Euro City Persons [1] (Pink), and Caltech Pedestrians [10] (Yellow). Y-axis values are % based in both (a, b). The proposed LSFM beats the human baseline on the Caltech dataset [47].

Recent research focuses on improving pedestrian detectors in terms of accuracy while ignoring their computational costs [5]. The performance of pedestrian detectors has improved a lot recently. However, there is still room for improvement, especially in heavy occlusion and small cases [14,15,20,24]. Fig. 1 shows the performance improvements in pedestrian detection over the last decade. Further, an improvement in accuracy usually comes with an increase in inference time [5], especially in the case of methods based on Vision-Transformers (ViT) [6, 9, 30]. A similar trade-off can be observed when using multi-modal sensor fusion. The accuracy improves while bearing heavy computational costs. A major component of ViT is self-attention, which has a complexity of $O(n^2)$ and does not scale well for high-resolution images [41]. Researchers have proposed alternatives to self-attention to avoid heavy computational costs, one of which is MLPMixers [37]. MLPMixer alternates between the channel and token dimension, thus maximizing cache efficiency, and achieving almost similar performance to transformers in image classification. However, when the image resolution is high, the MLPMixer feature map sizes increase quadratically, making it memory and compute-intensive backbones for downstream tasks. Also, the fully-connected nature of the MLP-based networks prevents them from being resolution independent like convolutions, as the number of parameters needs to be predefined.

We propose a novel pedestrian detection network that includes a *Multi Layer Perceptron* (MLP) based neck and a patched MLP mixer-based object detection head [37]. The proposed neck efficiently extracts and enriches key features from different stages of the backbone, and the detection head enables the dense connections between high-level semantic features. Together, when combined with a backbone, they constitute a lightweight, cache-efficient, and yet performant pedestrian detector. To train our network to be immune to motion blur and occlusion, we used hard mixup augmentation, which provides our network with data for soft occlusion and motion blur-like effects. Also, the hard mixup augmentation generates additional data for small detection cases to help the network absorb the key features which work across all scales.

We conduct an exhaustive evaluation of the proposed network on renowned pedestrian datasets to test it against the existing state-of-the-art methods in terms of both performance and efficiency. We conduct a cross dataset evaluation to test the domain generalization capabilities of the proposed network. Further, we perform the ablation study to check the effectiveness of different components of the proposed network. Major contributions of this work are as follows:

- We propose *Super Pixel Pyramid Pooling* (SP3), a MLP-based feature pyramid network.
- We propose *Dense Focal Detection Network* (DFDN), a lightweight head to allow denser connections.
- We pre-trained a deep but not wide ConvMLP [21] based backbone, ConvMLP Pin, for the proposed network to reduce inference time.
- We propose pedestrian detectors with backbones of different sizes to enable applications in resource-constrained environments.
- Our proposed model beats the human baseline [47] for the first time in the history of pedestrian detection.

## 2. Related Work

**RCNN Model Family:** Ross Girschick *et al*. proposed Region-based Convolutional Neural Networks (RCNNs) [13] as an early deep learning based solution to object detection using neural networks while utilizing selected search [40] to generate region proposals. Fast RCNN [12] proposed a single-stage pipeline that used the region of interest (RoI) pooling layer to share convolutions across all region proposals, hence sharing a lot of computation and decreasing inference times. Followed by Fast RCNN, Faster RCNN [34] enabled end-to-end training by proposing a novel deep learning based region proposal network to generate region proposals. He *et al*. proposed Mask R-CNN [16] as a powerful baseline system for instance segmentation, thereby improving the baseline for object detection based on Faster R-CNN. Cascade R-CNN [4] was proposed to address problems with degrading performance with increased Intersection over Union (IoU) thresholds. Cai *et al*. introduced Cascade Mask R-CNN, which extends Cascade R-CNN to instance segmentation by incorporating a mask head [4].

**Vision Transformers for Object Detection:** Since their introduction, Transformers [41] have gained popularity and found various applications in natural language processing and computer vision. Dosovitskiy *et al*. [11] proposed Vision Transformers (ViT) as a powerful alternative to convolution-based networks that reshapes images into patches for feature extraction. DEIT [39] introduces an attention-based distillation method along with data augmentation to improve performance without pretraining. DETR [6] makes use of transformers as a foundational block to handle object detection achieving better results compared to models such as Faster R-CNN [34] on the standard COCO object detection dataset. Most recently, UViT [7] was proposed as a single-scale Transformer for object detection, which omits the hierarchical pyramid designs used in earlier detectors and improves performance on COCO object detection as well as instance segmentation.

**MLP Mixer Based Architectures:** Tolstikhin *et al*. proposed MLP-Mixer [37], a non-convolutional object detection architecture that is solely built on multi-layer perceptrons (MLPs) applied over either spatial locations or feature channels. ResMLP [38] proposed a deeper architecture compared to MLP-Mixer while simplifying the token mixer and achieving better performance. gMLP [25] proposed a Spatial Gating Unit to process spatial features, improving the efficacy of the token-mixing MLPs.

**Anchor Free Pedestrian Detectors:** Anchor-free pedestrian detectors skip region proposal networks and directly predict pedestrians in a high-level semantic feature fashion using fully connected CNNs. CSP [29] detects pedestrians by predicting the center and scale map to reconstruct the bounding boxes. Adaptive center and scale prediction ACSP [43] makes use of switchable normalization during training on various batch sizes for better convergence and improved recall. F2DNet [20] improves the performance by introducing a second stage to the anchor-free detectors *i.e*., the fast suppression head.

## 3. Localized Semantic Feature Mixers

Inspired by their efficiency, we aim to develop a pedestrian detection model based on MLPMixers [37]. In order to enable our model to process variable-sized input, we use a ConvMLP-based backbone [21] and an MLPMixer-based detection head, which works with patches containing local information. To keep our network light, we avoid using a feature pyramid network and deploy novel, cache-efficient SP3. We use the center and scale representation of pedestrians which is considered a high-level semantic features representation [29] and therefore, call our network *Localized Semantic Feature Mixers*. Fig. 2 shows the detailed architecture of the components of LSFM. In the rest of the section, we will go through each component in detail.

### 3.1. Super Pixel Pyramid Pooling

*Feature Pyramid Network* (FPN) [22] enables the detection network to detect objects at different scales and fully utilize the different levels of features extracted from the backbone. The first step of FPN is to merge the feature maps from different backbone stages into a single, uniform-sized feature map array for further processing. Since concatenation is only possible for feature maps of the same spatial dimension, upscaling and downscaling operations are required to scale feature maps to a uniform size. Commonly used upscaling methods used in feature pyramid networks are transposed convolutions and interpolation [4, 29]. Both transposed convolution and interpolation have heavy computational and memory costs, although these layers do not contribute directly to the learning aspect of the network.

We propose *Super Pixel Pyramid Pooling* (SP3), a novel neck for our pedestrian detector, which takes a rather direct approach by applying a linear layer to filter and enrich features coming from the different stages of the backbone in a single operation. All feature maps of varying sizes are split into an equal number of patches with different resolutions. This is done by reducing patch size with the reducing size of feature-maps per stage *i.e*. $8 \times 8$ for the first stage, $4 \times 4$ for the second, and so on. Patches across the feature maps are then grouped based on the spatial location they correspond to, flattened in spatial and channel dimensions, and concatenated to form a single feature vector representing a spatial region across all stages of the backbone. We call this representation *Super Pixel Form*. These super pixels are then passed through a linear layer to achieve the desired number of features while filtering and enriching them. Finally, the resultant filtered super pixels get reshaped into patches for further processing. In this way, SP3 achieves its purpose of

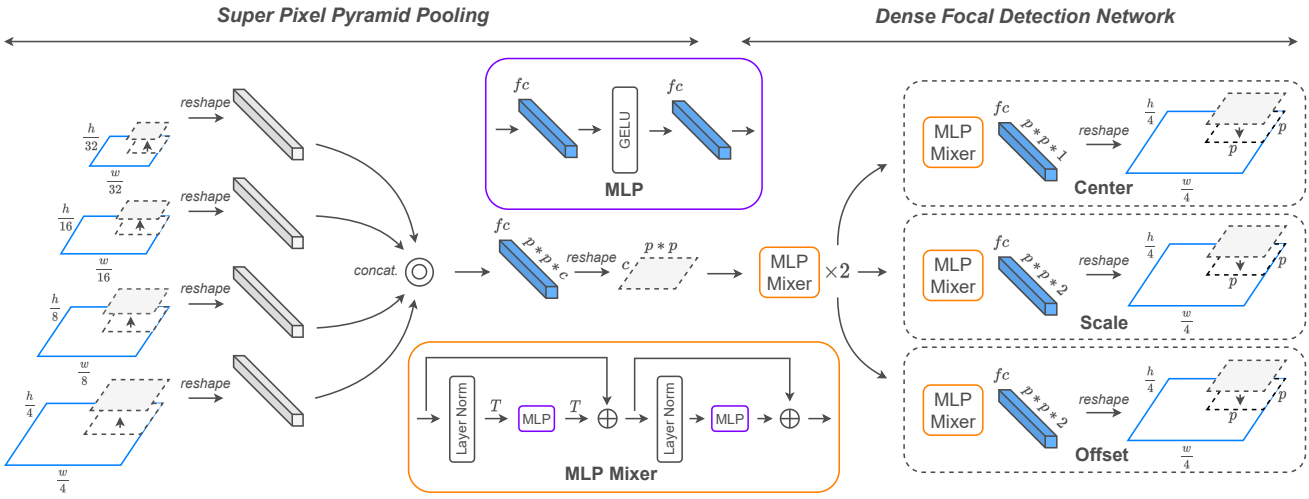



Figure 2. Shows the architecture of the *Super Pixel Pyramid Pooling* (SP3) neck followed by the *Dense Focal Detection Network* (DFDN), along with their components *i.e.*, MLP and MLP Mixer.

feature enrichment efficiently. The detailed architecture of SP3 is shown in Fig. 2.

### 3.2. Dense Focal Detection Network

The role of an object detection head is to convert the final feature embedding into objects. Since features received at the detection head have a larger spatial context, introducing further local spatial connections helps to extend it, which can further refine detections. In the detection heads of anchor-free approaches, detection attributes are predicted per pixel using convolutional layers. We propose the *Dense Focal Detection Network* (DFDN), a novel detection head entirely composed of MLPMixer layers [37]. The MLPMixer layers enable the efficient use of cache to boost the inference of the network. The DFDN works on patches instead of entire images and refines the detections based on local context information. This way, the complexity of MLPMixer layers becomes independent of input resolution resulting in scalability to higher resolutions.

Similar to [20], we use the center and scale representation of pedestrians, classify each pixel into the center of a pedestrian or not and regress the height and width of the pedestrian centered at that pixel. We follow the loss settings of the *Focal Detection Network* in F2DNet [20] and use offset maps for precise pedestrian centers. We use three MLPMixer blocks with 2 as the MLP expansion ratio in our DFDN to achieve better performance. Fig. 2 shows the detailed architecture of the DFDN. Further, we use $SmoothL1$ [12] loss for offset regression, $VanilaL1$ loss with log scaled height and width values for scale regression, and $\alpha$-balanced *Focal Loss* [23] for center prediction with $10^{-1}$, $5 \times 10^{-2}$ and $10^{-2}$ as their respective loss weights. The following shows the formulation of the center loss,

$$L_{center} = \frac{1}{K} \sum_{t} FL(p_t, y_t), \tag{1}$$

where

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} log(p_t) \quad , \qquad \alpha_t = \begin{cases} 1 & \text{if } y_t = 1 \\ (1 - M_t)^{\beta} & \text{otherwise.} \end{cases} \tag{2}$$

In the above-mentioned equation, $\alpha$ is the balancing or penalty reduction factor based on $M_t$ which is a gaussian kernel around the true positives. The values of $\beta$ and $\gamma$ are kept same as in [20] *i.e.*, $\beta = 4$ and $\gamma = 2$.

### 3.3. ConvMLP Backbone

ConvMLP proposed by Li *et al*. uses convolution layers in between MLP layers to enable spatial connections [21]. It is independent of the input resolution and requires comparatively low computational effort. Due to their linear memory footprint, MLP layers achieve high cache efficiency and have significantly higher inference speeds. Therefore, we choose ConvMLP [21] as a faster and more efficient backbone. Since we use semantic feature representations of pedestrians (the centers and scale representation), a deeper network is required to learn such complex functions precisely. Therefore, we design a deeper but not wide ConvMLP-based backbone and call it ConvMLP-Pin.

The first stage of ConvMLP-Pin contains a tokenizer and a residual bottleneck block to extract conventional features [17, 21]. The remaining three stages contain several ConvMLP blocks followed by downsampling at the end of each stage. We chose 4, 8, and 4 as the number of blocks

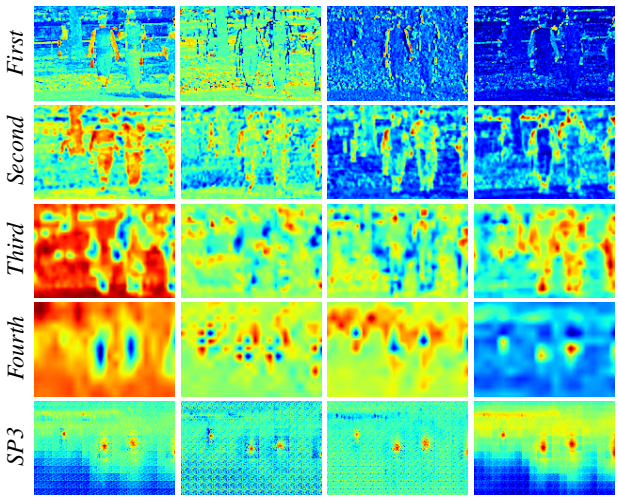



Figure 3. Feature maps of ConvMLP-Pin stages and SP3.

for the second, third and fourth stages respectively. With the abovementioned setting, the backbone becomes deeper; we set the MLP hidden dimension ratio to 2, to keep it light as well. We pretrain our ConvMLP-Pin backbone on the ImageNet-1000 dataset for 100 epochs and use it for pedestrian detection. Fig. 3 shows activation maps of the ConvMLP-Pin backbone and *Super Pixel Pyramid Pooling* (SP3). Where first four rows show the activation maps of different stages of the backbone and the last row shows the results of the SP3 layer, which combines features from all stages into single-sized, stacked feature maps.

### 3.4. Hard Mixup Augmentation

Autonomous driving datasets like [1, 10, 31, 48] were recorded using a camera mounted on a moving vehicle. Due to the setup, the images exhibit motion blur, which is hard to treat and hinders the training of deep learning models [15]. Also, a low number of heavily occluded cases gives the model only a glimpse of the occlusion phenomenon during training, but not enough to understand occlusions properly. Further object-aware augmentations like Cutmix [44] and Erase [49] can add undesired gradient artifacts to the image. Mixup augmentation [45] is an image-aware augmentation technique, widely used in training robust image classifiers. We applied a new variation of mixup augmentation for pedestrian detection by training the network with mixed-up samples and hard labels. Unlike mixup augmentation for classification, we do not use soft labels but keep all annotations with their original labels instead. Hard mixup augmentation provides the model with soft occluded samples for training and makes it robust to motion blur. We used mixup ratios in the range $(0.4, 0.6)$ so that all the objects still have enough information to be detected. However, going beyond this ratio requires the definition of a threshold that designates annotations to keep.

Table 1. Summary of the pedestrian detection datasets.

| Dataset | Images | Density | Time | Resolution |
|---|---|---|---|---|
| Caltech Ped. | 42,782 | 0.32 | day | 640 × 480 |
| City Persons | 2,975 | 6.47 | day | 2048 × 1024 |
| ECP | 21,795 | 9.2 | day, night | 1920 × 1024 |
| TJU-Ped-Traffic | 13,858 | 2.0 | day, night | 1624 × 1200 |

Table 2. Evaluation settings for pedestrian datasets.

| Setting | CP, Caltech, TJU-Traffic-Ped. | | Euro City Persons | |
|---|---|---|---|---|
| | Visibility | Height | Visibility | Height |
| Reasonable | [0.65, ∞] | [50, ∞] | [0.6, ∞] | [40, ∞] |
| Small | [0.65, ∞] | [50, 75] | [0.6, ∞] | [30, 60] |
| Heavy Occ. | [0.2, 0.65] | [50. ∞] | [0.2, 0.6] | [40. ∞] |
| All | [0.2, ∞] | [20. ∞] | [0.2, ∞] | [20. ∞] |

### 3.5. Mean Teacher Knowledge Distillation

Averaging the weights of the network during training results in a generalized network which is good for domain adaptation, as it prevents the network from overfitting [18]. Mean Teacher [36] takes the running mean of a network's checkpoints while training and saves them as teacher checkpoints. Unless indicated otherwise, we report the results of the proposed models based on mean teacher checkpoints.

## 4. Experimental Setup

This section contains the details of our experimental setup including datasets, evaluation metric, evaluation settings, and finally inference time calculation setup.

**Datasets:** Since the proposed model is specific to autonomous driving, only datasets containing traffic scenes were used for training. The recently published Euro City Persons dataset [1] contains 47,300 images encompassing scenes from 31 different cities of Europe. The dataset is vast, and captures different weather and lighting conditions [1]. The Euro City Persons dataset [1] contains both day and night scenes. However, only daytime scenes were used for training and testing in this work. The City Persons dataset [48] contains day scenes from 27 different cities in Germany. The image dimensions are almost the same as Euro City Persons images [1]. However, the City Persons [48] dataset is sparser compared to Euro City Persons [1] dataset. The Caltech Pedestrian dataset [10] has been used for a long time in pedestrian detection. It has much lower pedestrian density and low image resolution compared to the City Persons [48] and Euro City Persons datasets [1]. For training and evaluation on the Caltech dataset [10] we used the corrected annotation proposed by [47]. The TJU-DHD-Traffic [31] dataset contains traffic scenes with illumination and weather variance which increases the robust-

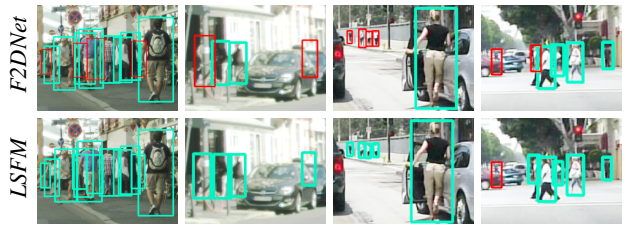



Figure 4. Qualitative comparison of LSFM and F2DNet [20]. Cyan marks the true positives and red marks the false negatives.

ness of pedestrian detectors when used for training. Tab. 1 shows details of all datasets used in this work.

**Evaluation Measure:** The evaluation measure used in this work is $MR^{-2}$. $MR^{-2}$ represents the area under the log average miss rate over the $fppi$ curve by taking the mean of miss rates at nine different $fppi$ thresholds equally divided in the log space from $(10^{-2}, 10^{0})$. All $MR^{-2}$ results reported in this work are $\%$ based.

**Evaluation Settings:** Different evaluation settings have been proposed in previous works to judge the performance of pedestrian detectors in different scenarios. In this work, the evaluation settings proposed with the Caltech pedestrian dataset [10] are used. These settings divide pedestrian detections into four overlapping subgroups based on the visibility ratio and pixel height of each detection, namely reasonable, small, heavy occlusion, and all. Visibility and height thresholds are different for the Euro City Persons dataset [1] while both City Persons and Caltech Pedestrian [10] datasets use the same thresholds proposed with the Caltech Pedestrian dataset [10]. Tab. 2 shows the different evaluation settings used in this work in detail. Unless mentioned otherwise, the results presented in this work are based on the evaluation of validation sets in the case of the City Persons [48], Euro City Person [1] and TJU-DHD-Pedestrian [31] datasets and the test set in case of the Caltech dataset [10].

**Inference Time Calculation:** To be consistent with existing pedestrian detectors on inference time calculation we used a GTX 1080Ti [20, 28, 35, 46]. The inference is done with a single image per batch on the original resolution.

## 5. Results

In this section, we discuss benchmarking results of LSFM. The listed results for LSFM are with two backbones *i.e.*, ConvMLP-Pin, HRNet [42]. LSFM indicates results with HRNet backbone [42] while LSFM P indicates results where ConvMLP-Pin was used. To save FLOPS and parameters, we only used the first three stages of HRNet since it contains fusion layers that aggregate information of the fourth stage to all other layers [42]. Tab. 3 shows benchmarking results of LSFM with previous state-of-the-art methods in a single dataset setting.

Table 3. Comparison with the state-of-the-art models shows that the LSFM performs significantly better in most settings while having the least inference times.

| Method | Reasonable | Small | Heavy | Inference |
|---|---|---|---|---|
| **City Persons** [48] | | | | |
| Pedestron [14] | 11.2 | 14.0 | 37.0 | 0.73s |
| CSP [14, 29] | 11.0 | 16.0 | 49.3 | 0.33s |
| PRNet [35] | 10.8 | - | 42.0 | 0.22s |
| APD [46] | 8.8 | - | 46.6 | 0.16s |
| F2DNet [20] | 8.7 | 11.3 | 32.6 | 0.44s |
| LSFM P (ours) | 8.7 | **8.7** | 32.4 | **0.13s** |
| LSFM (ours) | **8.5** | 8.8 | **31.9** | 0.18s |
| **Caltech** [10] | | | | |
| Pedestron [14] | 6.2 | 7.4 | 55.3 | 0.20s |
| ALFNet [28] | 6.1 | 7.9 | 51.0 | 0.05s |
| AR-Ped [2] | 4.4 | - | 48.8 | 0.09s |
| F2DNet [20] | **2.2** | **2.5** | 38.7 | 0.14s |
| LSFM P (ours) | 3.9 | 4.2 | 37.6 | **0.03s** |
| LSFM (ours) | 3.1 | 3.4 | **35.8** | 0.09s |
| **Euro City Persons** [1] | | | | |
| YOLOv3 [33] | 8.5 | 17.8 | 37.0 | - |
| FRCNN [34] | 7.3 | 16.6 | 52.0 | - |
| Pedestron [14] | 6.6 | 13.6 | 33.3 | 0.44s |
| F2DNet [20] | 6.1 | 10.7 | 28.2 | 0.41s |
| LSFM P (ours) | 7.0 | 13.5 | 30.0 | **0.13s** |
| LSFM (ours) | **4.7** | **9.9** | **23.8** | 0.17s |
| **TJU-Pedestrian-Traffic** [31] | | | | |
| F2DNet [20] | 21.6 | 26.3 | 62.6 | 0.40s |
| CrowdDet [8] | 20.8 | - | 61.2 | - |
| EGCL [24] | 19.7 | - | 60.1 | 0.76s |
| Pedestron [14] | 18.9 | **24.0** | 56.3 | 0.40s |
| LSFM P (ours) | 19.7 | 25.8 | 60.1 | **0.13s** |
| LSFM (ours) | **18.7** | 24.9 | **56.2** | 0.18s |

### 5.1. Qualitative Comparison

Fig. 4 shows a qualitative comparison of F2DNet [20] and LSFM on City Persons [48] and Caltech [10] datasets where images with diverging results are shown. It is evident that LSFM performs better especially in small and heavy occlusion cases however, there are a few cases of extreme occlusion where LSFM fails as well. Also, we found some rare cases where a pedestrian is detected by F2DNet [20] but missed by LSFM.

### 5.2. Inference Time

Tab. 3 shows single image inference times of different pedestrian detectors. LSFM successfully achieves the lowest inference time among other pedestrian detectors. On the Caltech Pedestrian dataset [10] LSFM P achieved 33 $ms$ inference time, resulting in 30 $fps$, which is considered real-time and it is almost $\frac{1}{4}$ of the inference time of the previous state-of-the-art F2DNet [20]. On average, LSFM P achieves $\sim$71$\%$ lesser inference time while LSFM achieves $\sim$55$\%$

Table 4. Cross dataset evaluation results. LSFM shows similar generalizability patterns compared to the state-of-the-art. All listed methods use HRNet [42] backbone.

| Method | Train | Test | Reasonable | Small | Heavy |
|---|---|---|---|---|---|
| CSP [14, 29] | ECP | CP | 11.5 | 16.6 | 38.2 |
| Pedestron [14] | ECP | CP | 10.9 | 11.4 | 40.9 |
| F2DNet | ECP | CP | 10.1 | 12.1 | **36.4** |
| **LSFM (ours)** | ECP | CP | **9.4** | **11.1** | 37.8 |
| **LSFM (ours)** | CP | Caltech | 11.7 | 15.6 | 37.4 |
| F2DNet | CP | Caltech | 11.3 | 13.7 | 32.6 |
| CSP [14, 29] | CP | Caltech | 10.1 | 13.3 | 34.4 |
| Pedestron [14] | CP | Caltech | **8.8** | **9.8** | **28.8** |
| F2DNet | ECP | Caltech | 16.9 | 21.5 | 41.3 |
| **LSFM (ours)** | ECP | Caltech | 13.1 | 16.3 | 33.1 |
| CSP [14, 29] | ECP | Caltech | 10.4 | 13.7 | 31.3 |
| Pedestron [14] | ECP | Caltech | **8.1** | **9.6** | **29.9** |
| CSP [14, 29] | CP | ECP | 19.6 | 51.0 | 56.4 |
| Pedestron [14] | CP | ECP | 17.4 | 40.5 | 49.3 |
| **LSFM (ours)** | CP | ECP | 17.0 | 42.1 | 49.6 |
| F2DNet | CP | ECP | **11.6** | **14.7** | **40.0** |

Table 5. Results of the progressive fine-tuning show that LSFM models beat the state-of-the-art across all datasets. TJU and CT indicate TJU-Traffic-Pedestrian [31] and Caltech [10] datasets. The presented results are based on models retrained from scratch.

| Method | Training Strategy | Reas. | Small | Heavy | Infe. |
|---|---|---|---|---|---|
| Pedestron [14] | TJU → ECP → **CP** | 8.9 | 10.6 | 29.6 | 0.73s |
| F2DNet [20] | TJU → ECP → **CP** | 6.8 | 9.0 | 26.0 | 0.44s |
| LSFM P (ours) | TJU → ECP → **CP** | 7.0 | 7.1 | 28.0 | **0.13s** |
| LSFM (ours) | TJU → ECP → **CP** | **6.7** | **6.7** | **23.5** | 0.18s |
| Pedestron [14] | ECP → **CT** | 2.6 | 2.8 | 24.4 | 0.20s |
| F2DNet [20] | ECP → **CT** | 1.2 | 1.4 | 19.6 | 0.14s |
| LSFM P (ours) | ECP → **CT** | 1.6 | 0.7 | 22.9 | **0.03s** |
| LSFM (ours) | ECP → **CT** | **1.0** | **0.2** | **19.5** | 0.09s |
| F2DNet [20] | TJU → **ECP** | 6.0 | 11.1 | 29.1 | 0.41s |
| Pedestron [14] | TJU → **ECP** | 4.7 | 10.2 | 24.7 | 0.44s |
| LSFM P (ours) | TJU → **ECP** | 5.5 | 11.6 | 26.0 | **0.13s** |
| LSFM (ours) | TJU → **ECP** | **4.1** | **9.5** | **20.9** | 0.17s |

lesser inference time, compared with F2DNet [20].

### 5.3. Comparison With The State-Of-The-Art

Tab. 3 compares LSFM with the state-of-the-art in single dataset settings. LSFM shows overall better performance compared to the state-of-the-art with on average $\sim 1\%$ $MR^{-2}$ reduction. However, LSFM P shows slightly higher $MR^{-2}$ compared to the state-of-the-art in most cases which, given its lowest inference time, is a better option for systems with limited resources. Overall, LSFM models perform well with superior performance compared to F2DNet [20] and present better tradeoffs.

Table 6. Comparison with human baseline shows LSFM beats human baseline on Caltech dataset [47].

| Method | Reasonable | Inference |
|---|---|---|
| Pedestron [14] | 1.75 | 0.20s |
| F2DNet [20] | 1.21 | 0.14s |
| Human Bl. [47] | 0.88 | - |
| LSFM (ours) | **0.87** | **0.09s** |

Table 7. Performance on the test set of City Persons. LSFM establishes a new state-of-the-art.

| Method | Reasonable | Small | Heavy | Inference |
|---|---|---|---|---|
| FRCNN [48] | 13.0 | 37.2 | 50.5 | - |
| Cascade R-CNN [4] | 11.6 | 13.6 | 47.1 | 0.73s |
| AdaptiveNMS [26] | 11.4 | 13.6 | 47.0 | - |
| MGAN [32] | 9.3 | 11.4 | 41.0 | - |
| APD-Pretrain [46] | 7.3 | 10.8 | 28.1 | - |
| Pedestron [15] | 7.7 | 9.2 | 27.1 | 0.73s |
| LSFM (ours) | **6.4** | **7.9** | **24.7** | **0.18s** |

Table 8. Performance on the test set of Euro City Persons. LSFM performs slightly inferior compared to SPNet [19]. * marks inference times calculated on Nvidia V100 GPU.

| Method | Reasonable | Small | Heavy | Inference |
|---|---|---|---|---|
| SSD [27] | 13.1 | 23.5 | 46.0 | - |
| Faster R-CNN [34] | 10.1 | 19.6 | 38.1 | - |
| YOLOv3 [33] | 9.7 | 18.6 | 40.1 | - |
| APD [46] | 5.3 | 12.4 | 26.8 | - |
| Pedestron [15] | 5.1 | 11.2 | 25.4 | 0.44s |
| LSFM (ours) | 4.4 | 10.6 | 22.9 | **0.17s** |
| SPNet [19] | **4.2** | **9.5** | **21.6** | 0.27s* |

### 5.4. Cross Dataset Generalization

Furthermore, we conduct a cross-dataset evaluation study to test the generalizability of LSFM models to unseen data. The experiments involve evaluating models on datasets other than the one used for training. In this way, the results of the cross-dataset evaluation give insights into the networks' ability to successfully transfer features learned on one dataset, onto another. Tab. 4 shows the cross-validation results of LSFM models in comparison with the state-of-the-art models. LSFM models show comparable cross-dataset validation results which prove that LSFM models generalize well to unseen data.

### 5.5. Progressive Fine-tuning

To test the performance of the LSFM on large datasets and how they scale with increasing dataset size, we conduct a progressive fine-tuning study. In progressive fine-tuning, the model is first trained on larger and more diverse datasets and gradually fine-tuned toward the target dataset. Tab. 5 shows detailed progressive fine-tuning results where $A \rightarrow B$ indicates training on dataset A followed by fine-tuning

Table 9. Ablation study on the components of LSFM and Hard Mixup Augmentation (H. Mix.).

| Backbone | FSH | H. Mix. | M. Tea. | SP3 | DFDN | Reasonable | Small | Heavy | Inference | Parameters | Flops |
|---|---|---|---|---|---|---|---|---|---|---|---|
| HRNet [42] | | | | | | 9.5 | 14.5 | 35.4 | 0.44s | 40.31M | 1774.3G |
| HRNet [42] | | | | | ✓ | 9.9 | 15.2 | 35.9 | 0.28s | 29.5M | 350.8G |
| HRNet [42] | | | | ✓ | ✓ | 9.5 | 13.5 | 33.9 | 0.18s | 32.5M | 347.1G |
| HRNet [42] | | | ✓ | ✓ | ✓ | 9.2 | 10.8 | 32.7 | 0.18s | 32.5M | 347.1G |
| HRNet [42] | | ✓ | ✓ | ✓ | ✓ | 8.8 | **8.6** | 32.7 | 0.18s | 32.5M | 347.1G |
| HRNet [42] | ✓ | ✓ | ✓ | ✓ | ✓ | **8.5** | 8.8 | **31.9** | 0.18s | 34.9M | 348.0G |
| ConvMLP-Pin | | | | ✓ | ✓ | 10.5 | 16.7 | 37.7 | 0.13s | 20.1M | 174.9G |
| ConvMLP-Pin | | | ✓ | ✓ | ✓ | 9.8 | 10.8 | 34.8 | 0.13s | 20.1M | 174.9G |
| ConvMLP-Pin | | ✓ | ✓ | ✓ | ✓ | **8.7** | 9.1 | 34.3 | 0.13s | 20.1M | 174.9G |
| ConvMLP-Pin | ✓ | ✓ | ✓ | ✓ | ✓ | **8.7** | **8.7** | **32.4** | 0.13s | 22.5M | 175.8G |

on B. The proposed LSFM models beat the state-of-the-art with a significant margin while taking significantly less time for inference in comparison. Tab. 6 shows a comparison of LSFM models with the human baseline on the Caltech dataset [47]. The proposed LSFM beats the human baseline for the first time in the history of pedestrian detection.

We also evaluate the proposed model on test sets of City Persons [48] and Euro City Persons [1]. Tabs. 7 and 8 show the performance of LSFM on the test set of the City Persons [48] and Euro City Persons datasets, respectively. The evaluation was done on the official servers of the datasets as annotations of these sets are withheld. LSFM outperforms previous state-of-the-art methods on the test set of City Persons [48], establishing a new state-of-the-art. However, the performance on the test set of Euro City Persons [1] is slightly inferior to the current state-of-the-art.

### 5.6. Ablation Study

In this section, the effects of individual parts of LSFM on the performance and inference time are examined. For the purpose of the ablation study, we used the City Persons dataset [48] only. Tab. 9 sums up the results of the ablation study. The study includes experiments on two different backbones *i.e.*, ConvMLP-Pin and HRNetW32 [42]. The first row of Tab. 9 shows the results of our baseline which is F2DNet without suppression head [20].

**Dense Focal Detection Network** (DFDN) is a lightweight detection head compared to *Focal Detection Head* [20]. The first two rows of Tab. 9 show pedestrian detection results of the model without DFDN and with DFDN, respectively. It is evident that using the DFDN reduces overall parameters and FLOPs resulting in lesser inference time with a slight drop in performance compared to the baseline.

**Super Pixel Pyramid Pooling** (SP3) enables efficient feature enrichment and pooling of features. Tab. 9 shows that although the number of parameters and FLOPs increase when using the SP3, the inference time still drops significantly along with notable performance improvements. This proves that SP3 is the more efficient, yet more performant alternative of *Feature Pyramid Network*.

**Mean Teacher** (M. Tea.) knowledge distillation keeps the running mean of a network's checkpoints to avoid overfitting. The resulting model is, therefore, more general and performant, as it is evident from Tab. 9.

**Hard Mixup Augmentation** (H. Mix.) provides LSFM with extra data and soft occlusions. The second last row for each backbone in Tab. 9 shows the results when using mixup augmentation. It is evident that the extra data is especially helpful in small pedestrian cases. Also, the artificial soft occlusions created by the hard mixup augmentation help LSFM to learn occlusions, ultimately resulting in better performance in heavily occluded cases.

**Fast Suppression Head** (FSH) [20] suppresses false positives, improving the quality of the final detections even further. Tab. 9 shows significant improvements in performance when using FSH, with a barely notable increase in inference time and model parameters.

## 6. Conclusion

This work presents a novel anchor-free pedestrian detection architecture with efficient components. The novel architecture uses MLPMixers and fully connected layers for denser connections and cache efficiency. This design achieves better performance with significantly reduced inference time. Further, the proposed models beat state-of-the-art pedestrian detectors on all mentioned datasets as well as the human baseline for the first time in the history of pedestrian detection. Since this work focuses on pedestrian detection in day scenes, an extension will be to study the scalability of the proposed models to night scenes. Also, it will be interesting to explore the potential of the proposed models to adapt to non-traffic scenarios, as this study solely focuses on pedestrian detection in traffic scenes.

## Acknowledgment

This work was funded by the German Ministry for Economic Affairs and Climate Action (BMWK) project KI Wissen under Grant 19A20020G. We would like to thank Adriano Lucieri for the useful discussions and insights.

## References


[1] Markus Braun, Sebastian Krebs, Fabian Flohr, and Dariu M Gavrila. Eurocity persons: A novel benchmark for person detection in traffic scenes. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):1844–1861, 2019. 2, 5, 6, 8

[2] Garrick Brazil and Xiaoming Liu. Pedestrian detection with autoregressive network phases. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7231–7240, 2019. 6

[3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[4] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: Delving into high quality object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6154–6162, 2018. 3, 7

[5] 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*, 2021. 1, 2

[6] 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*, pages 213–229. Springer, 2020. 2, 3

[7] Wuyang Chen, Xianzhi Du, Fan Yang, Lucas Beyer, Xiaohua Zhai, Tsung-Yi Lin, Huizhong Chen, Jing Li, Xiaodan Song, Zhangyang Wang, et al. A simple single-scale vision transformer for object localization and instance segmentation. *arXiv preprint arXiv:2112.09747*, 2021. 3

[8] Xuangeng Chu, Anlin Zheng, Xiangyu Zhang, and Jian Sun. Detection in crowded scenes: One proposal, multiple predictions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12214–12223, 2020. 6

[9] Xiyang Dai, Yinpeng Chen, Bin Xiao, Dongdong Chen, Mengchen Liu, Lu Yuan, and Lei Zhang. Dynamic head: Unifying object detection heads with attentions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7373–7382, 2021. 2

[10] 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. 2, 5, 6, 7

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 3

[12] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 3, 4

[13] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014. 3

[14] 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, 2021. 1, 2, 6, 7

[15] Irtiza Hasan, Shengcai Liao, Jinpeng Li, Saad Ullah Akram, and Ling Shao. Pedestrian detection: Domain generalization, cnns, transformers and beyond. *arXiv preprint arXiv:2201.03176*, 2022. 1, 2, 5, 7

[16] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[17] 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, 2016. 4

[18] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 5

[19] Chenhan Jiang, Hang Xu, Wei Zhang, Xiaodan Liang, and Zhenguo Li. Sp-nas: Serial-to-parallel backbone search for object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11863–11872, 2020. 7

[20] Abdul Hannan Khan, Mohsin Munir, Ludger van Elst, and Andreas Dengel. F2dnet: Fast focal detection network for pedestrian detection. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4658–4664. IEEE, 2022. 1, 2, 3, 4, 6, 7, 8

[21] Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. Convmlp: Hierarchical convolutional mlps for vision. *arXiv preprint arXiv:2109.04454*, 2021. 2, 3, 4

[22] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 3

[23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 4

[24] Zebin Lin, Wenjie Pei, Fanglin Chen, David Zhang, and Guangming Lu. Pedestrian detection by exemplar-guided contrastive learning. *IEEE transactions on image processing*, 2022. 2, 6

[25] Hanxiao Liu, Zihang Dai, David So, and Quoc V Le. Pay attention to mlps. *Advances in Neural Information Processing Systems*, 34:9204–9215, 2021. 3

[26] Songtao Liu, Di Huang, and Yunhong Wang. Adaptive nms: Refining pedestrian detection in a crowd. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6459–6468, 2019. 7

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European conference on computer vision*, pages 21–37. Springer, 2016. 7

[28] Wei Liu, Shengcai Liao, and Weidong Hu. Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding. *IEEE transactions on image processing*, 29:1413–1425, 2019. 6

[29] 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, 2019. 3, 6, 7

[30] 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, 2021. 2

[31] Yanwei Pang, Jiale Cao, Yazhao Li, Jin Xie, Hanqing Sun, and Jinfeng Gong. Tju-dhd: A diverse high-resolution dataset for object detection. *IEEE Transactions on Image Processing*, 30:207–219, 2020. 5, 6, 7

[32] 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, 2019. 7

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 6, 7

[34] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 6, 7

[35] 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*, pages 32–48. Springer, 2020. 6

[36] 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. 5

[37] Ilya O Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, et al. Mlp-mixer: An all-mlp architecture for vision. *Advances in Neural Information Processing Systems*, 34:24261–24272, 2021. 2, 3, 4

[38] Hugo Touvron, Piotr Bojanowski, Mathilde Caron, Matthieu Cord, Alaaeldin El-Nouby, Edouard Grave, Gautier Izacard, Armand Joulin, Gabriel Synnaeve, Jakob Verbeek, et al. Resmlp: Feedforward networks for image classification with data-efficient training. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3

[39] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. 3

[40] Jasper RR Uijlings, Koen EA Van De Sande, Theo Gevers, and Arnold WM Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013. 3

[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2, 3

[42] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10):3349–3364, 2020. 6, 7, 8

[43] Wenhao Wang. Adapted center and scale prediction: more stable and more accurate. *arXiv preprint arXiv:2002.09053*, 2020. 3

[44] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 5

[45] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 5

[46] Jialiang Zhang, Lixiang Lin, Jianke Zhu, Yang Li, Yun-chen Chen, Yao Hu, and Steven CH Hoi. Attribute-aware pedestrian detection in a crowd. *IEEE Transactions on Multimedia*, 23:3085–3097, 2020. 6, 7

[47] 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, 2016. 2, 5, 7, 8

[48] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3221, 2017. 1, 2, 5, 6, 7, 8

[49] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 13001–13008, 2020. 5