F2DNet: Fast Focal Detection Network for Pedestrian Detection

Abdul Hannan Khan*†, Mohsin Munir†, Ludger van Elst† and Andreas Dengel*†,
*Fachbereich Informatik, Technische Universität Kaiserslautern,
67663 Kaiserslautern, Germany
†German Research Center for Artificial Intelligence (DFKI GmbH),
67663 Kaiserslautern, Germany
Corresponding Author: hannan.khan@dfki.de

***Abstract*—Two-stage detectors are state-of-the-art in object detection as well as pedestrian detection. However, the current two-stage detectors are inefficient as they do bounding box regression in multiple steps i.e. in region proposal networks and bounding box heads. Also, the anchor-based region proposal networks are computationally expensive to train. We propose F2DNet, a novel two-stage detection architecture which eliminates redundancy of current two-stage detectors by replacing the region proposal network with our focal detection network and bounding box head with our fast suppression head. We benchmark F2DNet on top pedestrian detection datasets, thoroughly compare it against the existing state-of-the-art detectors and conduct cross dataset evaluation to test the generalizability of our model to unseen data. Our F2DNet achieves 8.7%, 2.2%, and 6.1% MR^{-2} on City Persons, Caltech Pedestrian, and Euro City Person datasets respectively when trained on a single dataset and reaches 20.4% and 26.2% MR^{-2} in heavy occlusion setting of Caltech Pedestrian and City Persons datasets when using progressive fine-tunning. On top of that F2DNet have significantly lesser inference time compared to the current state-of-the-art. Code and trained models will be available at https://github.com/AbdulHannanKhan/F2DNet.**

I. Introduction

Pedestrian detection is a sub-domain of object detection where the target class is pedestrian and the rest is considered background. Pedestrian detection plays a vital role in autonomous driving as well as surveillance. In autonomous driving, one of the most important objectives is to avoid collision with pedestrians by detecting and tracking them. This objective is to be carried out in a limited resource scenario as limited computational power is available inside an autonomous vehicle due to compactness and power efficiency constraints. This requires the pedestrian detection model to be light and efficient. Also, the lesser the time model takes to process a single frame more frame per second it can process which yields better awareness of surroundings.

Region Proposal Networks were first proposed by Ross Girshick et al. [1] to replace, slow, selective search-based region proposal generation with a faster, CNN-based network that can be trained end-to-end along with detection head. In the last decade, researchers have focused on improving two-stage detectors by proposing new detection heads [2], [3], [4] with little focus on region proposal network architecture. However, the role of region proposal networks in two-stage detectors

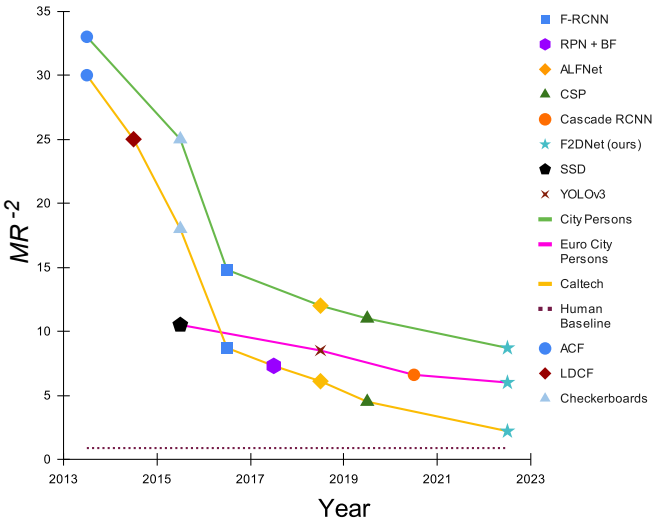


Fig. 1: Shows evolution of pedestrian detectors over the years and their corresponding MR^{-2} on Caltech Pedestrian [5], City Persons [6] and Euro City Persons [7] datasets in reasonable settings.

is limited to proposing candidate regions with the purpose of objectness score produced by region proposal networks limited to proposal filtering. Also, proposed bounding boxes from region proposal network need rigorous refinement in their coordinates, for example, Cascade RCNN [4] applies three cascading heads to get refined detections.

Compared to two-stage detectors, single-stage detectors are efficient as they split the image into a grid and perform detection per patch eliminating region proposal network [8], [9], [10], [11]. However, single-stage detectors do not perform as good as two-stage detectors in terms of accuracy, this can be attributed to a class imbalance between positive and negative samples [10]. Other than class imbalance, since each patch does not necessarily contain a full object, classifying if a patch contains enough parts of an object is sub-optimal, as a part may belong to multiple object classes. A common attribute of both single and two-stage detectors explained above is anchors. Both kinds of detectors rely on anchors with predefined aspect

ratios.

In the last few years, anchor-free object detectors were proposed [12], [13]. Motivated from anchor-free approaches in object detection, anchor-free approaches were proposed for pedestrian detection as well [14], [15], [16]. These pedestrian-specific approaches, take the idea of single-stage detector to another level by predicting classes per pixel instead of per patch. However, this is done on a downscaled feature map to be efficient and robust [14]. Unlike anchor-based approaches, center and scale-based approaches classify if each pixel is a center pixel of an object and regress the possible scale of that object. In this way, center and scale-based approaches eliminate the idea of predicting rough bounding boxes and refining them later on. Further, center and scale-based approaches use the focal loss as classification loss to deal with class imbalance [14].

Although center and scale-based approaches have optimal design and better convergence they produce more false positives due to penalty reduced focal loss, which does not punish much the false predictions in the neighborhood of positive pixels. This problem intensifies in the case of small and heavily occluded pedestrians.

Our method is different in nature from existing single and two-stage detectors. The closest single-stage detector is anchor-free, center and scale prediction [14]. However, we only use the head from CSP [14] with different loss settings as the CSP head [14] is stronger and efficient compared to region proposal networks, also we use fast suppression head to further refine detections. Compared to two-stage detectors, we replace the region proposal network with a stronger detection network, we do not call it another region proposal network because focal detection network produces strong detection candidates compared to proposals that need further bounding box refinement and classification. Also, we replace computationally expensive traditional second stage, which predicts bounding boxes as well as classifies them, with a simple and efficient suppression head to only suppress false positives without altering bounding boxes.

Contribution of this paper is three fold;

- First, we redesign a two-stage detection architecture to remove redundant and inefficient bounding box prediction and replace region proposal network with a strong detection network, followed by a light-weight suppression head instead of multiple bounding box heads.
- Second, we propose focal detection network as our classification and bounding box regression head, which can independently produce satisfactory results.
- Third, we propose fast suppression head to handle false positives produced by focal detection head in small and heavily occluded settings.

II. Related Work

Significant improvements have been made in recent years in the field of pedestrian detection using deep learning models [17], [2], [11], [14] as shown in Fig. 1. Most of the recent techniques follow general object detection workflow including a strong pre-trained backbone to extract features, an optional feature pyramid network (FPN) [18] based feature enrichment layer, a region proposal network (RPN) [1] in case of two-stage detectors and at the end, bounding box heads for bounding box regression and classification. Such pipelines are supported in modern object detection frameworks like mmdetection [19]. Different types of pedestrian detectors have emerged in recent years which can be differentiated from each other based on how they use region proposal network and choice of bounding box heads.

A. Anchor Based Pedestrian Detectors

Region-based convolutional neural networks are two-staged object detectors, which were first proposed by Girshick et al. in [20] for object detection. Fast-RCNN and Faster-RCNN were proposed to improve the processing time of RCNN by using ROI pooling on features maps instead of raw image and CNN-based region proposal network respectively [21], [1]. Mask Guided Attention Network incorporates additional visibility information of the object to handle occlusions better [2]. Cascade R-CNN proposed by Cai and Vasconcelos in [4] uses multiple bounding box heads to refine detections in cascading manner. Another anchor-based but single-stage pedestrian detector is ALFNet, which is based on Single Shot Multibox Detector (SSD) [11], [8]. RetinaNet is yet another single-stage detector, which is similar to SSD [8] but introduces focal loss to handle foreground and background class imbalance [22].

B. Anchor Free Pedestrian Detectors

Anchor-free pedestrian detectors are pedestrian-specific object detectors that do not use anchors or region proposal networks. Instead, they predict bounding box and class per pixel on down-scaled feature maps because performing detection per pixel on original resolution is costly. CornetNet [13] uses CNN based approach to predict paired keypoint heatmaps i.e. one heatmap for each top-left and bottom-right corner. Fully convolutional single-stage object detection network, FCOS was proposed in [12], which adopts classification and bounding box prediction of R-CNN heads to pixel-wise fashion with bounding box predictions being pixel distances from object center, which is calculated using centeredness predicted per pixel by FCOS. Center and Scale Prediction CSP [14] proposed for pedestrian detection uses a similar approach but instead predicts center heatmap, scale map and reconstructs the bounding boxes using the center and scale [14]. Adaptive center and scale prediction ACSP [15] uses switchable normalization for better convergence on different batch sizes and uses full resolution for training to improve recall. APD [16] tries to handle crowded pedestrians by additionally predicting density and diversity. BGCNet replaces normal convolutions with box-guided convolution for center heatmap subnet to incorporate predicted scale and offset information in center heatmap prediction [23].

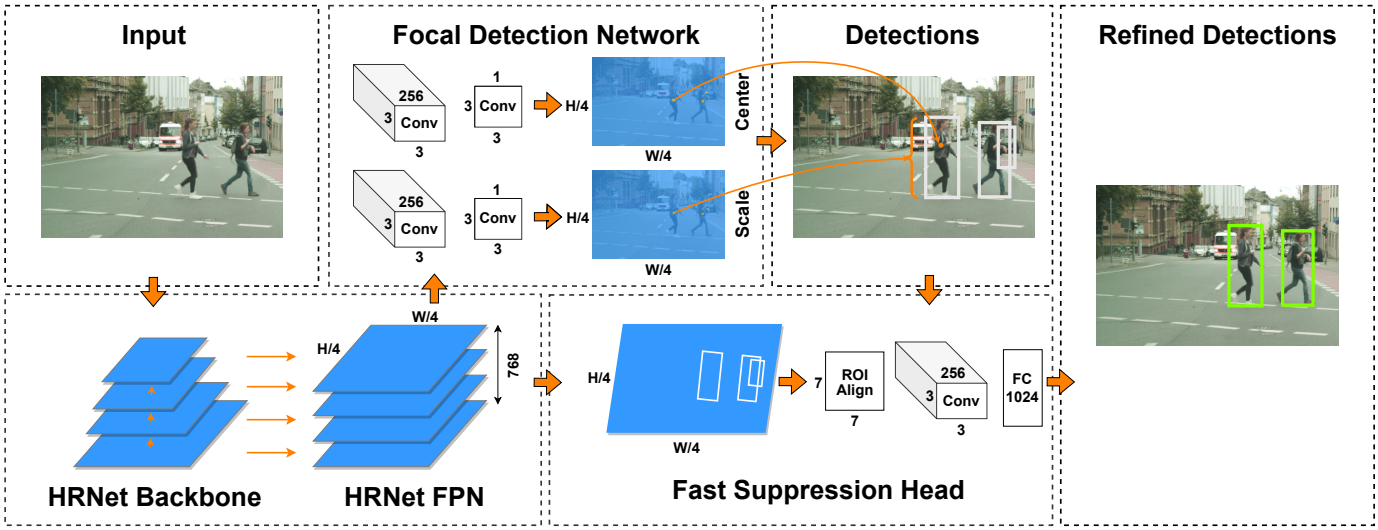


Fig. 2: The network architecture of our F2DNet. The input image is passed through backbone and FPN to extract feature maps which are then passed to the focal detection network to obtain initial detections. The detected bounding boxes are then passed to the fast suppressed head along with feature maps to suppress false positives.

C. ViT Based Methods

Vision transformer (ViT) is the adaptation of transformers in the domain of computer vision. DETR [24] and DETR for pedestrian detection [25] use ViT based bounding box heads to provide a wider receptive field. Recently proposed soft teacher based approach for object detection [26] uses SWin transformer [27] based backbone and semi-supervised training to produce promising results.

III. Fast Focal Detection Network

In this section, we explain the architecture of our fast focal detection network in detail and argue about our design choices. First, we elaborate on the feature extraction process followed by the detailed architecture of F2DNet and conclude the section by explaining the detection formation strategy. Fig. 2 shows complete architecture of our model.

A. Feature Extraction

To predict precise location and size high-resolution features are required which contain semantic and position information. Aggressive down and upscaling can result in loss of this vital information [17]. Therefore, we use the HRNetW32v2 backbone [28] for feature extraction as it extracts high-resolution features from images. To obtain feature maps of a single scale, we take feature maps from different stages of backbone, upscale them to $(h/4, w/4)$ using bilinear interpolation and apply convolution operations. In this way, the model stays light on memory as interpolation operation has no memory cost but is effective as succeeding convolution operations provide necessary learnable parameters.

B. Fast Focal Detection Network

Current two-stage object detection architectures employ a weak region proposal network followed by strong bounding box heads. We take a different approach and use a strong detection head succeeded by a light suppression head. In this way, the detection head focuses on precise localization and high classification recall while the suppression head takes care of false positives. In short, our two-stage detection architecture attains high efficiency by eliminating the repetition contained in current two-stage architectures. Architecture of our focal detection network and fast suppression head is detailed below.

1) Focal Detection Network: The architecture of the focal detection network is based on the idea of center and scale map prediction which eliminates explicit modeling of bounding boxes for detection [14]. Our approach is somewhat similar to that proposed in [14] however, we use different loss settings to fine-tune the architecture for better convergence and precise localization.

Center loss for focal detection network can be formulated as:

$$L_{center} = \frac{1}{K} \sum_{i} \sum_{j} \alpha_{ij}\, CE(p_{ij}, y_{ij}), \tag{1}$$

where

$$CE(p_{ij}, y_{ij}) = \begin{cases} -\log(p_{ij}) & \text{if } y_{ij} = 1 \\ -\log(1 - p_{ij}) & \text{otherwise,} \end{cases}$$
$$\alpha_{ij} = \begin{cases} (1 - p_{ij})^{\gamma} & \text{if } y_{ij} = 1 \\ p_{ij}^{\gamma}(1 - M_{ij})^{\beta} & \text{otherwise.} \end{cases} \tag{2}$$

In equation above, p_{ij} and y_{ij} are predicted center probability and ground truth label respectively. $CE(p_{ij}, y_{ij})$ represents cross entropy loss with α_{ij} being weight at each location (i, j). M_{ij} represents gaussian based penalty reduction for surrounding pixels of true centers as designation of exact center brings difficulty in training [14]. The p_{ij}^{γ} and $(1-p_{ij})^{\gamma}$

terms define focus weight based on prediction confidence i.e. it reduces contribution of easy examples to the loss and helps optimizer to focus on hard examples. The $(1 - M_{ij})^{\beta}$ term reduces loss for false positives closer to true centers. We used $\gamma = 2$ and $\beta = 4$ in our experiments.

In [21] Smooth L1 loss is recommended for regression as it is robust to outliers. The Smooth L1 loss reduces penalty when the distance between predicted and actual height is small, which helps in better convergence. However, since we use $\log$ of height instead of actual height value it can cause smaller detections and ultimately result in false positives due to insufficient IoU. Therefore, we use Vanilla L1 Loss as regression loss to make height predictions more accurate.

We define loss for the focal detection head as:

$$L_{FDN} = \lambda_r L_{reg} + \lambda_c L_{cls} + \lambda_o L_{off} \tag{3}$$

Where λ_r, λ_c and λ_o represent weights for regression, classification and offset loss respectively. We experimentally found $\lambda_r = 0.05$, $\lambda_c = 0.01$ and $\lambda_o = 0.1$ help model converge better than other weight settings.

2) Fast Suppression Head: Since, the focal detection network uses penalty-reduced focal loss as a center loss, false positives in the neighborhood of positive centers are not punished sufficiently. While most of these false positives are suppressed by Non-Maximum Suppression (NMS), it still needs another suppression step to suppress the rests i.e. where IoU with positive predictions is lower than 0.5. Therefore, we propose a simple and fast suppression head to further refine the detections. The simple architecture of the fast suppression head can be seen in Fig. 2. We train the fast suppression head in detached settings, i.e. the gradients from the fast suppression head do not flow back to feature maps or detection head. In this way, a simple, light yet effective suppression head is achieved. We use binary cross entropy as loss for our fast suppression head.

C. Pedestrian Detection

Each prediction gets one score from the focal detection network and another from the fast suppression head. We eliminate thresholding hyperparameter by combining both scores using the generative model shown in Fig. 3. We are particularly interested in an event where pedestrian is detected an not suppressed i.e. $P(\neg s, d|c, h)$. The detection model is derived from joint probability distribution of $P(s, d, c, h)$ and represented by following relation:

$$P(\neg s, d|c, h) = P(\neg s|d, c, h)P(d|c) \tag{4}$$

TABLE I: Details of pedestrian detection datasets.

Dataset	Images	Pedestrians	Density	Resolution
Caltech Pedestrians	42,782	13,674	0.32	640 × 480
City Persons	2,975	19,238	6.47	2048 × 1024
Euro City Persons	21,795	201,323	9.2	1920 × 1024

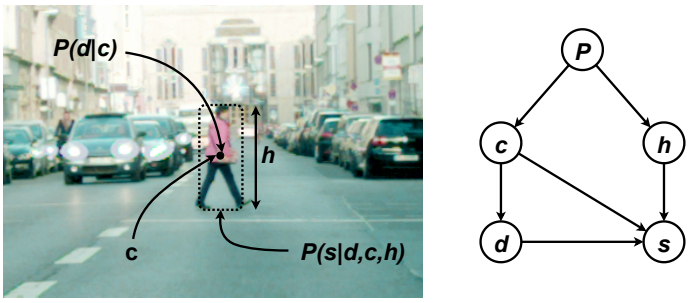


Fig. 3: (Left): Representation of pedestrian showing: the center c, height h, prediction from focal detection network $P(d|c)$ and prediction from fast suppression head $P(s|d, c, h)$. (Right): graphical representation of our pedestrian detection generative model where P represents pedestrian.

where c and h are the center and height of pedestrian respectively. $P(d|c)$ is the probability of a position detected as pedestrian center by focal detection head and $P(s|d, c, h)$ is the probability that given a bounding box detection it is suppressed by fast suppression head.

IV. Experimental Setup

In this section, our experimental setup is detailed, which we follow in the rest of the paper unless stated otherwise. First, we briefly go through datasets, followed by the evaluation settings in which we evaluate results on these datasets, and finally, we explain the evaluation criteria we used to compare our models with the existing state-of-the-art.

A. Datasets

To benchmark, our model we present results on three commonly used pedestrian detection datasets i.e. City Persons [6], Euro City Persons [7] and Caltech Pedestrian dataset [5]. Detailed statistics of these datasets can be seen in Table I.

All the results presented in this paper for Caltech pedestrian dataset [5] are based on its test set, while for City Persons [6] and Euro City Persons [7] results are based on their respective validation sets, unless stated otherwise.

B. Evaluation Settings

In pedestrian detection, evaluation settings define different subsets of a dataset which are used to better judge the performance of a model in different scenarios. We use evaluation settings proposed in Caltech Pedestrian [5] and Euro City Persons [7] datasets. Based on visibility and height of annotations these evaluation settings form four groups where each annotation can belong to more than one group. Settings followed across the paper can be seen in the Table II. It is important to note that evaluation settings are different for Euro City Persons dataset [7] while City Persons [6] and Caltech Pedestrian datasets [5] share identical evaluation settings.

TABLE II: Evaluation settings for pedestrian datasets based on height and visibility.

Setting	City Persons, Caltech		Euro City Persons	
	Visibility	Height	Visibility	Height
Reasonable	[0.65, ∞]	[50, ∞]	[0.6, ∞]	[40, ∞]
Small	[0.65, ∞]	[50, 75]	[0.6, ∞]	[30, 60]
Heavy Occlusion	[0.2, 0.65]	[50. ∞]	[0.2, 0.6]	[40. ∞]
All	[0.2, ∞]	[20. ∞]	[0.2, ∞]	[20. ∞]

C. Evaluation Criteria

We use Log-average miss rate over false positive per image or MR^{-2} to compare our model against recent models as it has been suggested in pedestrian detection datasets [6], [5], [7] as well as followed by the state of the art [17], [23], [3], [14]. $MR{-}2$ is calculated by taking geometric mean of miss rates at 9 equally spaced $ffpi$ thresholds in log space i.e. $fppi \in \{10^{-2}, 10^{-1.75}, ..., 10^{0}\}$.

D. Weighted Averaging

We used the mean teacher strategy of weighted averaging for better convergence and performance, as the model obtained after the weighted averaging performs better [29], [14]. All results of our models provided in this paper are based on the evaluation of the averaged model unless stated otherwise.

E. Training Details

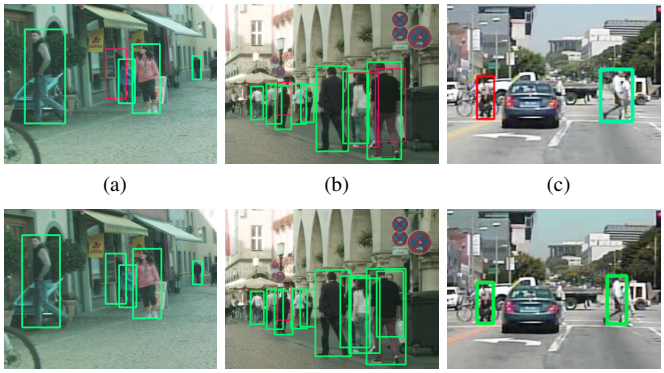
We used the Nvidia RTXA6000 GPU cluster to train our models. We used Distributed Data-Parallel to achieve parallel training on multiple GPUs with a manual seed. We used 2 GPUs with 32 and 4 images per GPU for training model on Caltech Pedestrian [5] and City Persons [6] datasets respectively. However, for training the model on Euro City Persons dataset [7] we used 4 GPUs with 4 images per GPU. We used a constant learning rate throughout the training after warm-up iterations with a maximum of 80 epochs.

V. Experiments and Results

In this section, we present the comparison of F2DNet to the current state-of-the-art and top-performing detectors based on MR^{-2} and inference time, along with the performance gains achieved by the suppression head.

Fig. 4 shows qualitative comparison of current state-of-the-art Cascade R-CNN [4] with our F2DNet. It shows that our F2DNet can detect pedestrians even where Cascade R-CNN fails. For results shown in Fig. 4, we took models trained on multiple datasets; for Cascade R-CNN we took model weights from Pedestron[17]. Fig. 5 first row shows that F2DNet without suppression head produces false positives most of which are suppressed by employing fast suppression head as shown in Fig. 5 second row.

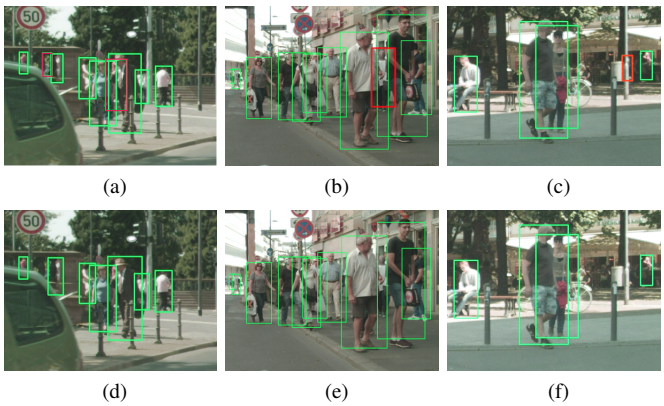
To compare the performance of F2DNet with the current state-of-the-art and top-performing methods, we took models trained on a single dataset without using any extra data except for pre-trained backbones. F2DNet outperforms the existing state-of-the-art in Caltech pedestrian [5] and Euro City Persons [7] datasets as well as in heavy occlusion settings of City Persons dataset [6] with a clear margin and achieves slightly better MR^{-2} in reasonable and small settings of City Persons dataset [6].



(a) (b) (c) (d) (e) (f)

Fig. 4: Qualitative comparison of F2DNet and Cascade R-CNN results. (a, b, c) Results of Cascade R-CNN on City Perons [6] and Caltech Pedestrians [5] datasets. Bounding Boxes marked red indicate false negatives. (d, e, f) Results of F2DNet on City Persons [6] and Caltech Pedestrian [5] datasets.

However, F2DNet with suppression head performs slightly worse compared to F2DNet without suppression head in heavy occlusion setting of Caltech pedestrian dataset [5]. This performance drop can be attributed to the sparseness of the Caltech pedestrian dataset [5] compared to other pedestrian datasets with less heavy occlusion samples to train suppression head well. Table III shows the detailed results our experiment.



(a) (b) (c) (d) (e) (f)

Fig. 5: Results of F2DNet before and after suppression. (a, b, c) Results without suppression; it can be seen that there are few false positives (marked red) in small and heavy occlusion cases. (d, e, f) Results with suppression; the false positives have been successfully suppressed.

TABLE III: Comparison of F2DNet with the current state-of-the-art detectors based on MR^{-2} and inference Time.

Method\MR^{-2}	Reasonable	Small	Heavy Occ.	Time
City Persons [6]				
ALFNet [11]	12.0	19.0	51.9	0.27s
Cascade R-CNN [17]	11.2	14.0	37.0	0.73s
CSP [14]	11.0	16.0	49.3	0.33s
PRNet [30]	10.8	-	42.0	0.22s
Beta R-CNN [3]	10.6	-	47.1	-
MGAN [2]	10.5	-	39.4	-
Adaptive CSP [15]	10.0	12.9	36.8	-
F2DNet no sup. (ours)	9.0	11.5	33.8	0.43s
BGCNet [23]	8.8	11.6	43.9	0.16s
F2DNet (ours)	**8.7**	**11.3**	**32.6**	0.44s
Caltech [5]				
Cascade R-CNN [17]	6.2	7.4	55.3	0.20s
ALFNet [11]	6.1	7.9	51.0	0.05s
CSP [14]	5.0	6.8	46.6	-
Rep Loss [31]	5.0	5.2	47.9	-
F2DNet no sup. (ours)	2.3	2.7	**38.2**	0.13s
F2DNet (ours)	**2.2**	**2.5**	38.7	0.14s
Euro City Persons [7]				
SSD [7]	10.5	20.5	42.0	-
YOLOv3 [7]	8.5	17.8	37.0	-
Faster R-CNN [7]	7.3	16.6	52.0	-
F2DNet no sup. (ours)	7.2	12.8	31.6	0.40s
Cascade RCNN [17]	6.6	13.6	33.3	0.44s
F2DNet (ours)	**6.1**	**10.7**	**28.2**	0.41s

TABLE IV: Cross dataset evaluation results. Our model is more generaliable compared to CSP and Cascade RCNN is most cases specially when trained on City Persons [6] and tested on Euro City Persons [7].

Method\MR^{-2}	Training	Reasonable	Small	Heavy Occ.
City Persons [6]				
CSP [17], [14]	ECP	11.5	16.6	38.2
Cascade RCNN [17]	ECP	10.9	**11.4**	40.9
F2DNet (ours)	ECP	**10.1**	12.1	**36.4**
Caltech [5]				
F2DNet (ours)	CP	11.3	13.7	32.6
CSP [17], [14]	CP	10.1	13.3	34.4
Cascade R-CNN [17]	CP	**8.8**	**9.8**	**28.8**
Euro City Persons [7]				
CSP [17], [14]	CP	19.6	51.0	56.4
Cascade RCNN [17]	CP	17.4	40.5	49.3
F2DNet (ours)	CP	**11.6**	**14.7**	**40.0**

VI. Cross Dataset Evaluation

We conduct cross dataset evaluation to test how well F2DNet generalizes to unseen data. We compare the generalizability of F2DNet, with two other models, namely CSP [14] and Cascade RCNN [4]. Both of these models are state of the art in the context of pedestrian detection. We used scores for Cascade RCNN [4] and CSP [14] provided in Pedestron [17]. We train F2DNet only on training sets and conduct tests on the validation set for City Persons [6] and Euro City Persons [7] and on the test set of Caltech Pedestrian dataset [5]. F2DNet generalizes better than CSP [14] and Cascade RCNN [4], in most cases, for City Persons [6] or Euro City Persons [7] (refer to Table IV). However, for Caltech dataset [5] F2DNet generalizes slightly worse than other models. F2DNet beats CSP [14] and Cascade RCNN [4] with a large margin when trained on City Persons [6] and tested on Euro City Persons [7], this shows that F2DNet performs well even when trained on a smaller dataset. Cross dataset evaluation scores can be seen in Table IV.

TABLE V: Results of F2DNet trained on multiple datasets in progression fine-tuning fashion.

Training	Testing	Reasonable	Small	Heavy Occ.
ECP $\rightarrow$ CP	CP	**7.80**	**9.43**	**26.23**
ECP $\rightarrow$ CP $\rightarrow$ Caltech	Caltech	**1.71**	**2.10**	**20.42**

VII. Progressive Fine Tuning

To further improve the performance of F2DNet we perform progressive fine-tuning. We initially train our model on a bigger and diverse dataset and fine-tune it towards the target dataset in cascading manner. For City Persons dataset [6], we train the model on Euro City Persons [7] and fine-tune on City Persons dataset [6]. For Caltech pedestrian dataset [5] we take the fine-tuned model on City Persons dataset [6] and fine-tune it on the Caltech pedestrian dataset [5]. Through progressive fine-tuning, we were able to achieve new all times low MR^{-2} in heavy occlusion settings for Caltech Pedestrian [5] and City Persons datasets [6] as shown in Table V. For both training and fine-tuning only train sets of respective datasets were used.

VIII. Conclusion

Two-stage detectors perform well in pedestrian detection however, the region proposal network-based two-stage detectors are inefficient as the region proposal networks are trained to predict weak proposals which need further refinement. We replaced the region proposal network with a strong focal detection network which is based on the per-pixel center and scale regression and hence produce high-quality candidates which, standalone are good detection except for some false positives in small and occluded settings. We pass these strong candidates from the focal detection network through a lightweight fast suppression network, which with barely noticeable computational cost further refines the detections to produce promising results. Our model beats state of the arts in most visibility and height settings while being on par in rest, without using any extra data except for pre-trained backbone. Also, by using Euro City Persons [7] and City Persons [6] datasets as extra training data, our model achieves the lowest MR^{-2} in a heavy occluded setting, in a multi-dataset setup.

REFERENCES

[1] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks." in *NIPS*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds., 2015, pp. 91–99. [Online]. Available: http://dblp.uni-trier.de/db/conf/nips/nips2015.html#RenHGS15

[2] Y. Pang, J. Xie, M. H. Khan, R. Anwer, F. Khan, and L. Shao, "Mask-guided attention network for occluded pedestrian detection," 10 2019, pp. 4966–4974.

[3] Z. Xu, B. Li, Y. Yuan, and A. Dang, "Beta r-cnn: Looking into pedestrian detection from another perspective," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran Associates, Inc., pp. 19 953–19 963.

[4] Z. Cai and N. Vasconcelos, "Cascade r-cnn: Delving into high quality object detection," 12 2017.

[5] P. Dollár, C. Wojek, B. Schiele, and P. Perona, "Pedestrian detection: An evaluation of the state of the art," *PAMI*, vol. 34, 2012.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," 06 2016.

[7] M. Braun, S. Krebs, F. B. Flohr, and D. M. Gavrila, "Eurocity persons: A novel benchmark for person detection in traffic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[8] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "Ssd: Single shot multibox detector," 12 2015.

[9] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," 06 2016, pp. 779–788.

[10] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, 07 2018.

[11] W. Liu, S. Liao, and W. Hu, "Efficient single-stage pedestrian detector by asymptotic localization fitting and multi-scale context encoding," *IEEE Transactions on Image Processing*, vol. PP, pp. 1–1, 09 2019.

[12] Z. Tian, C. Shen, H. Chen, and H. Tong, "Fcos: Fully convolutional one-stage object detection," 10 2019, pp. 9626–9635.

[13] H. Law and J. Deng, "Cornernet: Detecting objects as paired keypoints," *International Journal of Computer Vision*, vol. 128, 03 2020.

[14] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu, "High-level semantic feature detection: A new perspective for pedestrian detection," 06 2019, pp. 5182–5191.

[15] W. Wang, "Adapted center and scale prediction: More stable and more accurate," 02 2020.

[16] J. Zhang, L. Lin, Y.-C. Chen, S. Hoi, and J. Zhu, "Csid: Center, scale, identity and density-aware pedestrian detection in a crowd," 10 2019.

[17] I. Hasan, S. Liao, J. Li, S. Ullah Akram, and L. Shao, "Pedestrian detection: The elephant in the room," 03 2020.

[18] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," 07 2017, pp. 936–944.

[19] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

[20] A. Virasova, D. Klimov, O. Khromov, I. Gubaidullin, and V. Oreshko, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Radioengineering*, pp. 115–126, 01 2021.

[21] R. Girshick, "Fast r-cnn," 04 2015.

[22] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, 07 2018.

[23] J. Li, S. Liao, H. Jiang, and L. Shao, "Box guided convolution for pedestrian detection," 10 2020, pp. 1615–1624.

[24] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, *End-to-End Object Detection with Transformers*, 11 2020, pp. 213–229.

[25] M. Lin, C. Li, X. Bu, M. Sun, C. Lin, J. Yan, W. Ouyang, and Z. Deng, "Detr for pedestrian detection," 12 2020.

[26] M. Xu, Z. Zhang, H. Hu, J. Wang, L. Wang, F. Wei, X. Bai, and Z. Liu, "End-to-end semi-supervised object detection with soft teacher," *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[28] J. Wang, S. ke, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, pp. 1–1, 04 2020.

[29] A. Tarvainen and H. Valpola, "Weight-averaged consistency targets improve semi-supervised deep learning results," 03 2017.

[30] X. Song, K. Zhao, W.-S. Chu, H. Zhang, and J. Guo, *Progressive Refinement Network for Occluded Pedestrian Detection*, 11 2020, pp. 32–48.

[31] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen, "Repulsion loss: Detecting pedestrians in a crowd," 06 2018, pp. 7774–7783.