Self-Mimic Learning for Small-scale Pedestrian Detection -Supplementary Material-

A EXTENSION TO ONE-STAGE DETECTORS

To validate the generalization ability of our proposed SML, we extend it to one-stage object detectors. Specifically, we build SML on top of a state-of-the-art anchor-free pedestrian detector CSP [2]. As shown in Figure 1, the CSP has two branches placed on top of the feature maps generated by the feature extractor. The top branch predicts a center heatmap for the input image, where each location in the feature map has a confidence score indicating how likely the location is the center of a pedestrian. This branch can be considered as a classifier which determines the centers of pedestrians in the image. The bottom branch predicts a scale map from which the height of a candidate pedestrian centered at each location can be derived. The CSP is a fully convolutional network and no RoI pooling layer is used in it. To apply SML to the CSP detector, we insert two transform layers, one RoI align layer and one mimic loss layer into the CSP as shown in Figure 1. Each transform layer is a 3×3 convolutional layer with 128 output channels. The transform layer in the top branch outputs new feature maps on which the proposed mimic loss is imposed (see Section 4 for more details on self-mimic learning). The mimic loss serves to enhance the features of small-scale pedestrians with the help of the features of large-scale pedestrians in the new feature maps, such that better center confidences could be predicted for small-scale pedestrians. Since proposals are not available in one-stage detectors, we use ground truth boxes to which the RoI Align operation is applied (see Figure 1). In our experiments, we find that the performance is unstable if we directly apply SML to the last feature maps (the green feature maps in Figure 1) of the CSP detector. This is because that the scale prediction branch (bottom branch) is scale-aware, *i.e.,* predicting larger values for large-scale pedestrians and small values for small-scale pedestrians. The SML could make the features less discriminative (*i.e.,* the features of small-scale and large-scale pedestrians become similar) for scale prediction if it is applied to both branches. To address this issue, we introduce the two transform layers and only apply the mimic loss to the center prediction branch such that the center heatmap and scale map are predicted from different feature maps.

We experiment our approach on the CityPersons [3] dataset. Following [2], we train the model for 150 epochs on the training set with the same hyper-parameters and report the best results from epochs 50 to 150 in Table 1. We study two methods of applying our SML to CSP. In the first method, we directly apply SML to the last feature maps (green feature maps in Figure 1), and we denote this method by CSP+SML. In the second method, we insert the two transform layers and apply SML to the new classification feature maps (purple feature maps in Figure 8), and we denote this method by CSP*+SML. As shown in Table 1, compared with the CSP, our methods CSP+SML and CSP*+SML significantly improve the performance by a large margin of 1.8% and 3.3% MR^{-2} on the Small subset, respectively. We also experiment the CSP detector with the transformation module (CSP*). The CSP* performs worse than the

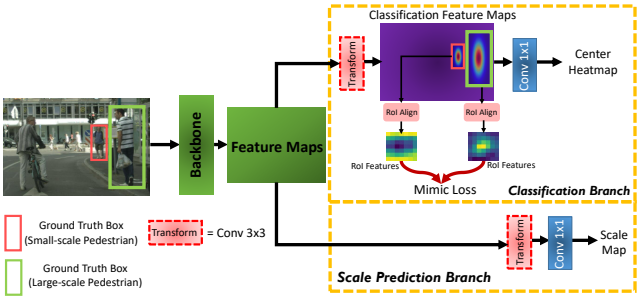

Figure 1: Extension of SML to the one-stage detector CSP [2].

Table 1: Results of the one-stage detector CSP [2] and our approach on the CityPersons dataset. $*$ denotes the CSP equipped with the two transform layers. Following [2], the best miss rates from epochs 50 to 150 are reported.

Method	Reasonable	Small
CSP	11.6	17.4
CSP*	14.2	19.3
CSP+SML	11.7	15.6
CSP*+SML	**10.6**	**14.1**

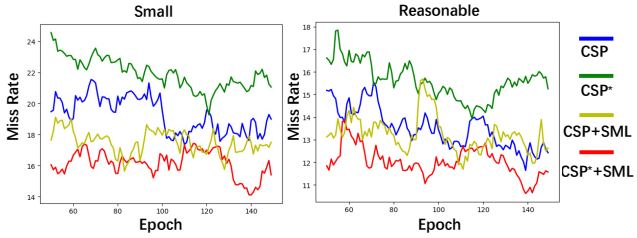

Figure 2: Results of the one-stage detector CSP [2] and our approach on the CityPersons dataset. $*$ denotes the CSP equipped with the two transform layers. Following [2], we train the CSP detector for 150 epochs and evaluate the performance from the epoch 50 to epoch 150.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413634

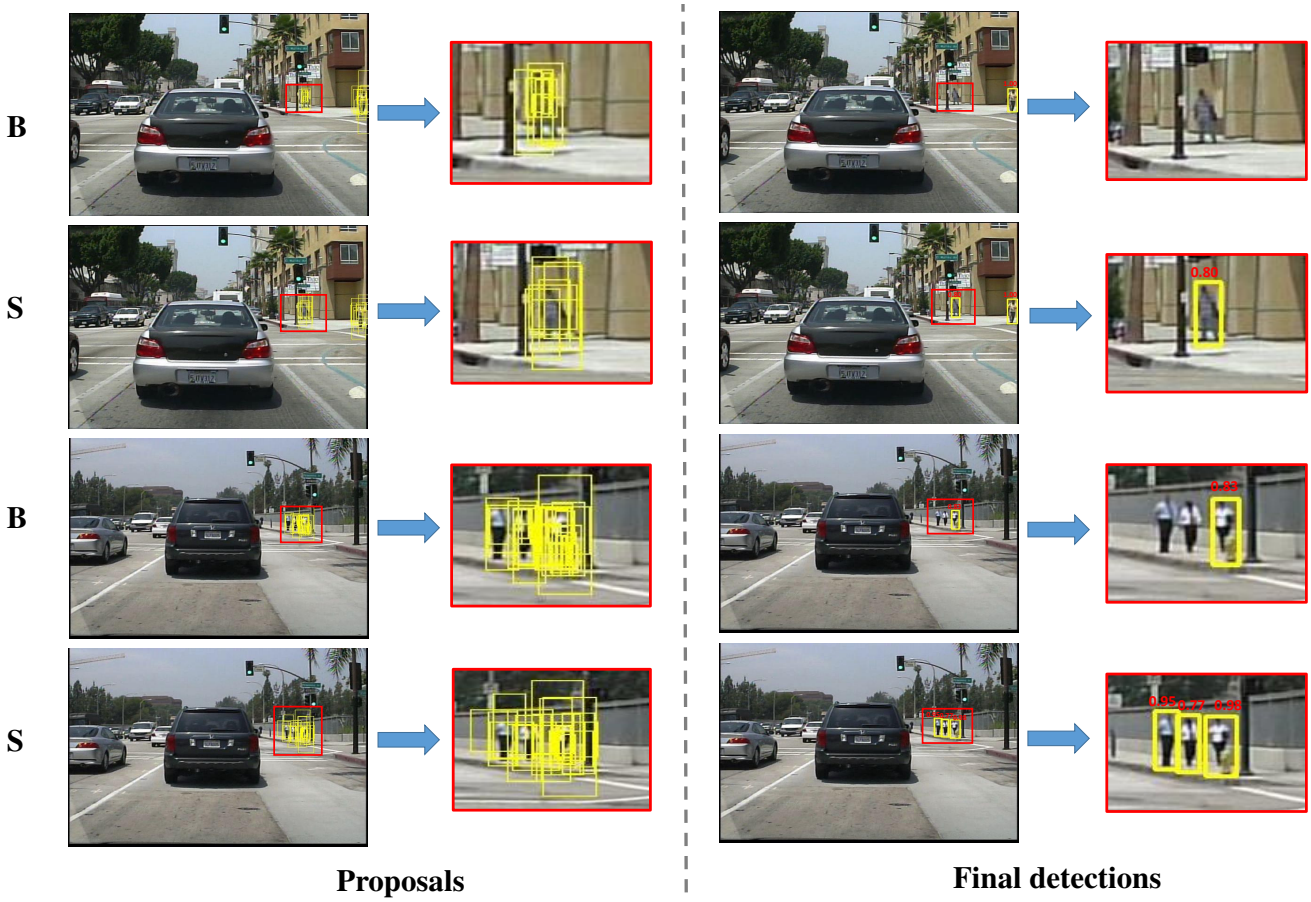

Figure 3: Qualitative comparison of the baseline and SML on the Caltech dataset. B denotes the baseline and S denotes our proposed SML. Examples on the left of the dotted line show proposals generated by the RPN (after NMS and filtering proposals with low scores), and examples on the right of the dotted line show final detection results (after removing boxes whose classification scores are lower than 0.5.). Red boxes denote areas where the baseline detector cannot correctly classify tiny pedestrians.

original CSP detector, while the results of CSP*+SML are better than those of CSP+SML. This is probably because the convolutional layer with kernel size of 1×1 in the original CSP detector is sufficient for center heatmap estimation, and the introduced transform layer in the classification branch may cause overfitting for the CSP*. In contrast, the SML can serve as an effective regularizer to avoid such overfitting for the classification branch. Moreover, it suggests that the performance gain of CSP*+SML comes from the proposed SML instead of the additional network layers. Figure 2 shows the miss rate curves on the CityPersons validation set from epoch 50 to epoch 150. We observe that the performance of CSP+SML is relatively unstable during the 100th epoch as we point out above. Notably, the miss rate curves of CSP*+SML are consistently better than those of the CSP and CSP* on both the Reasonable and Small subsets. These results demonstrate that the proposed SML is effective to further improve the performance for CSP on small-scale pedestrians. More importantly, it validates that our proposed SML can serve as a general component which can be incorporated into both one-stage and two-stage detectors to further improve the performance especially on small-scale pedestrians.

B QUALITATIVE RESULTS

Caltech. The qualitative examples on the Caltech [1] dataset are shown in Figure 3. As we can see from the examples, both the baseline method and our proposed SML can well locate tiny pedestrians by their RPNs. However, the baseline method miss them in second stage due to the weak feature representations of small-scale pedestrians. In contrast, with the help of self-mimic learning, SML compensates the missing details of small-scale pedestrians in feature space and enhances their feature representations, making it easier to classify them.

CityPersons. As shown in Figure 4, compared to the baseline detector, SML obtains more discriminative features for the small-scale pedestrians, which helps the detection head network to better

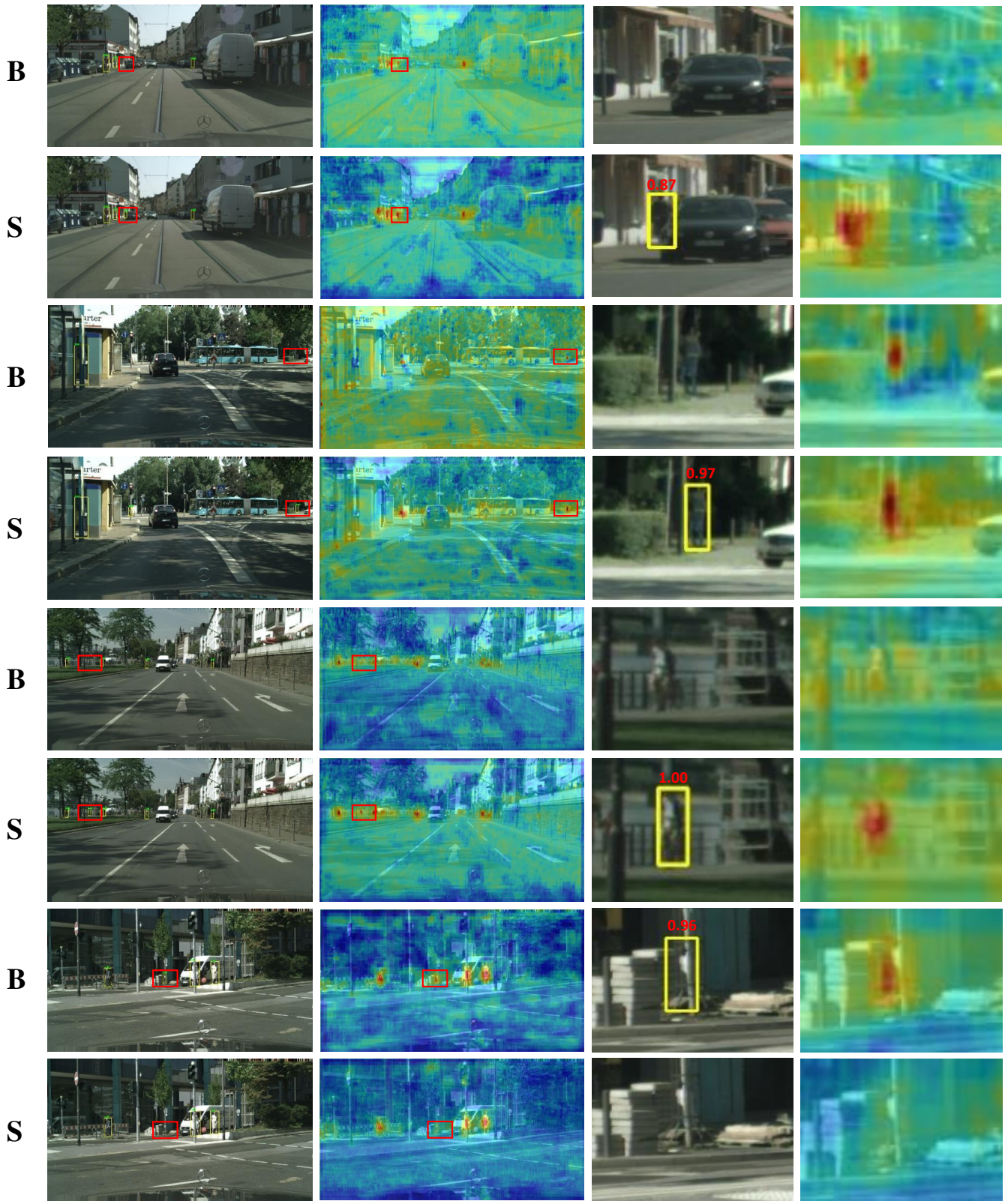

Figure 4: Qualitative comparison of the baseline and SML on the CityPersons dataset. B denotes the baseline and S denotes our proposed SML. The enlarged regions are selected from the red boxes in original images, and the feature maps are selected from the P2 in FPN. We remove boxes whose classification scores are lower than 0.5.

Table 2: Results of SML with different height thresholds.

Dataset	Threshold	H_S (H_L=120)				H_L (H_S=80)		
		60	70	80	90	100	140	160
Caltech	Reasonable	6.7	**6.6**	6.8	7.4	6.7	6.9	7.1
	Medium	21.8	21.6	**21.2**	21.8	21.8	21.5	22.1
	Far	63.9	**62.9**	63.6	63.9	63.4	63.6	63.8
CityPerson	Reasonable	12.7	12.5	**12.3**	12.5	12.4	12.4	12.4
	Small	19.8	19.5	19.3	19.2	19.8	19.1	**18.7**

Table 3: Results of SML with different α.

Dataset	α	1	4	8	16	32	64
Caltech	Reasonable	7.0	7.0	7.1	6.8	6.8	**6.4**
	Medium	22.8	22.2	21.6	**21.2**	21.6	21.6
	Far	64.9	65.0	64.1	63.6	62.1	**60.7**
CityPersons	Reasonable	12.9	**12.3**	12.5	**12.3**	12.9	12.8
	Small	20.7	19.7	18.5	19.3	**18.4**	19.8

distinguish them from backgrounds. Moreover, our method also effectively suppresses the false positives as shown in the last example of Figure 4. These qualitative examples demonstrate strong ability of SML for detecting very tiny pedestrians.

C HYPER-PARAMETERS.

Height Threshold. We experiment with different settings of H_S and H_L. The results are shown in Table 2. We can see from the results that our approach is not sensitive to H_S and H_L and outperforms the baseline detector consistently under different settings of height thresholds.

Mimic Loss Weight. We experiment SML with different mimic loss weight α in the Table 3. Our approach has relatively low performance with $\alpha < 4$ and the performance becomes consistently better with $4 \leq \alpha \leq 64$.

REFERENCES

[1] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona. 2012. Pedestrian detection: An evaluation of the state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 743–761.

[2] Wei Liu, Shengcai Liao, Weiqiang Ren, Weidong Hu, and Yinan Yu. 2019. High-level semantic feature detection: a new perspective for pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5187–5196.

[3] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. 2017. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3213–3221.