Temporal-Context Enhanced Detection of Heavily Occluded Pedestrians - Supplementary Material -

1. Qualitative Evaluation of Heavily Occluded Pedestrian Detection

As shown in Fig. 1, we visualize some proposal tubes and detection results on the Caltech dataset [2]. Compared to the baseline detector and the baseline+FGFA [1], our proposed method achieves better detection of heavily occluded pedestrians.

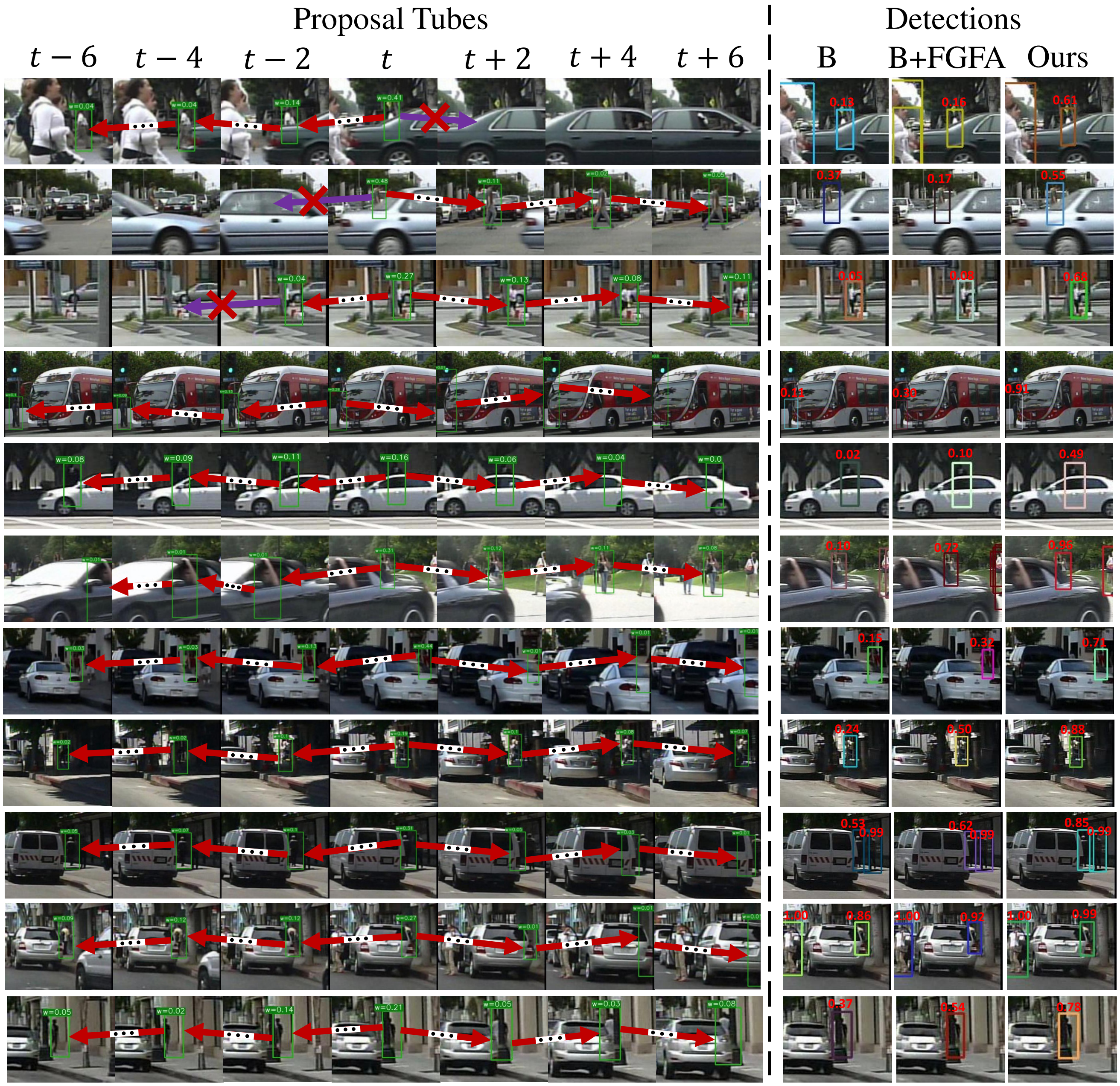

Figure 1. Qualitative comparison with the baseline (B) and the baseline+FGFA (B+FGFA) on the Caltech dataset. Ours indicates the TFAN+TDEM+PRM. w in the figures denotes the adaptive weight. The purple arrow indicates that the tube linking procedure is terminated as there is no proposal in adjacent frame which has an $IoU > \varepsilon$ with the reference proposal. For clear visualization, only one tube is shown in each row.

2. Qualitative Evaluation of False Alarm Suppressing

For pedestrian detection especially in night time, many ambiguous negative samples, *e.g.,* trees and poles, are usually misclassified with a high confidence score by a single-frame detector. To qualitatively validate the effectiveness of our method for false alarm suppressing, we visualize some proposal tubes and detection results on the NightOwls dataset [3] as shown in Fig. 2. By exploiting local temporal context, our method is able to access more references from neighboring frames, which can help the classifier more confidently suppress the hard negative examples.

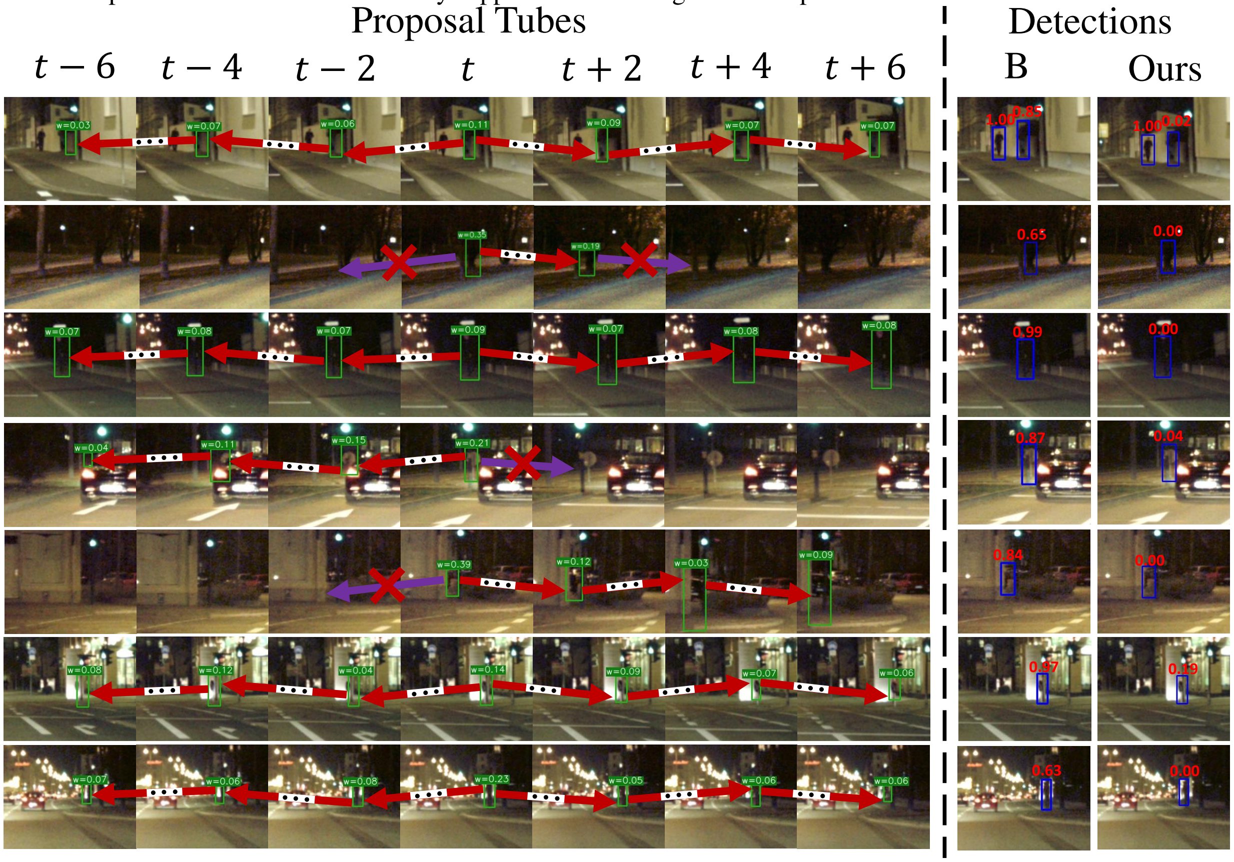

Figure 2. Qualitative comparison with the baseline (B) on the NightOwls dataset. Ours indicates the TFAN+TDEM. w in the figures denotes the adaptive weight. The purple arrow indicates that the tube linking procedure is terminated as there is no proposal in adjacent frame which has an $IoU > \varepsilon$ with the reference proposal. For clear visualization, only one tube is shown in each row.

3. Hyper-parameters

We experiment the proposed TFAN (+TDEM+PRM) with different settings of hyper-parameters, and the tube length τ is set to 6 by default. λ is used for enlarging the gap among examples when calculating the adaptive weights. As shown in Table 1, the performance becomes better when λ is set to larger than 1, and $\lambda = 5$ is a suitable choice in experiments. σ is the parameter for measuring the relative location similarity between two proposals. We can observe from Table 1 that the TFAN is not sensitive to σ. Table 2 shows the ablation study of the TFAN with different γ, which is utilized for the PRM module to retain sufficient pixels for measuring the semantic similarity of embedding features. The γ is set to a value such that at least $\gamma\%$ pixels in the embedding features are retained. We can see that the performance of the TFAN is stable when γ is set to $20\% \leqslant \gamma \leqslant 50\%$. As the γ becomes larger, it leads to a worse result, since excess pixels may introduce more background features for heavily occluded pedestrians.

λ	1	2	3	4	5	6	7	8	9	10	σ	0.2	0.3	0.4	0.5	0.6	0.7	0.8
R+HO	13.6	12.8	12.6	12.5	12.4	12.4	12.8	12.9	13.1	13.2	R+HO	12.8	12.7	12.6	12.4	12.6	12.6	12.6
HO	33.6	32.1	31.6	31.5	30.9	31.3	32.1	32.5	33.0	33.2	HO	31.9	31.9	31.5	30.9	31.9	32.1	32.1
R	7.5	7.0	6.8	6.7	6.7	6.8	7.1	7.1	7.1	7.2	R	6.8	6.8	6.7	6.7	6.7	6.8	6.8

Table 1. Ablation study of the TFAN with different λ and σ on the Caltech dataset.

γ	20	35	50	75	90	100
R+HO	12.4	12.4	12.4	12.6	12.6	12.9
HO	30.9	30.9	31.0	32.0	32.0	32.7
R	6.7	6.7	6.7	6.8	6.9	6.8

Table 2. Ablation study of the TFAN with different γ on the Caltech dataset.

References

[1] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei. Flow-guided feature aggregation for video object detection. In *ICCV*, 2017. 1

[2] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. In *IEEE TPAMI*, 2012. 1

[3] L. Neumann, M. Karg, S. Zhang, C. Scharfenberger, E. Piegert, S. Mistr, O. Prokofyeva, R. Thiel, A. Vedaldi, A. Zisserman, and B. Schiele. Nightowls: a pedestrians at night dataset. In *ACCV*, 2018. 2