Temporal-Context Enhanced Detection of Heavily Occluded Pedestrians

Jialian Wu[1] Chunluan Zhou[2] Ming Yang[3] Qian Zhang[3] Yuan Li[3] Junsong Yuan[1]
[1]State University of New York at Buffalo [2]Wormpex AI Research [3]Horizon Robotics, Inc.
{jialianw,jsyuan}@buffalo.edu czhou002@e.ntu.edu.sg
m-yang4@u.northwestern.edu {qian01.zhang,yuanli}@horizon.ai

Abstract

State-of-the-art pedestrian detectors have performed promisingly on non-occluded pedestrians, yet they are still confronted by heavy occlusions. Although many previous works have attempted to alleviate the pedestrian occlusion issue, most of them rest on still images. In this paper, we exploit the local temporal context of pedestrians in videos and propose a tube feature aggregation network (TFAN) aiming at enhancing pedestrian detectors against severe occlusions. Specifically, for an occluded pedestrian in the current frame, we iteratively search for its relevant counterparts along temporal axis to form a tube. Then, features from the tube are aggregated according to an adaptive weight to enhance the feature representations of the occluded pedestrian. Furthermore, we devise a temporally discriminative embedding module (TDEM) and a part-based relation module (PRM), respectively, which adapts our approach to better handle tube drifting and heavy occlusions. Extensive experiments are conducted on three datasets, Caltech, NightOwls and KAIST, showing that our proposed method is significantly effective for heavily occluded pedestrian detection. Moreover, we achieve the state-of-the-art performance on the Caltech and NightOwls datasets.

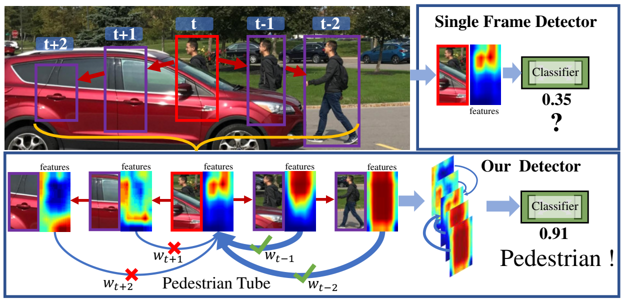

Figure 1. Top row: A heavily occluded pedestrian often leads to miss detection for a single frame detector due to incomplete and weak observations. Bottom row: In our approach, we exploit local temporal context of a heavily occluded pedestrian, *i.e.*, similar less occluded pedestrians in nearby frames, to enhance its feature representations. After linking temporally these pedestrian samples to a tube, we aggregate their features using an adaptive weight scheme by matching between visible parts, which substantially help to distinguish the heavily occluded one from the background.

1. Introduction

Detecting heavily occluded pedestrians is crucial for real-world applications, *e.g.*, autonomous driving systems, and remains the Gordian Knot to most state-of-the-art pedestrian detectors [27, 28, 10, 26, 24, 23, 19, 17, 54, 47, 46, 15, 16]. This challenge boils down to two aspects: **(i)** Heavily occluded pedestrians are hard to be distinguished from background due to missing/incomplete observations; **(ii)** Detectors seldom have a clue about how to focus on the visible parts of partially occluded pedestrians. Many great efforts have been made to address the occlusion issue, *e.g.*, attention mechanisms [29, 9], feature transformation [11] and part-based detection [22, 19, 13]. While these occlusion handling approaches alleviate partially occluded pedestrian detection in still images, they may not bring extra information beyond a single image for detectors to reliably infer an occluded pedestrian in essence. In this paper, we argue that the temporal context can essentially enhance the discriminability of the features of heavily occluded pedestrians which has not been studied thoroughly in previous works.

Our key idea is to search for non/less-occluded pedestrian examples (which we call them *reliable* pedestrians) with discriminative features along temporal axis, and if they are present, to exploit them to compensate the missing information of the heavily occluded ones in the current frame, as shown in Fig. 1. Specifically, our approach is carried out with two main steps. **(i)** *Tube linking*: starting from a pedestrian proposal in the current frame, we iteratively search for its relevant counterparts (not necessarily the same person) in adjacent frames to form a tube; **(ii)** *Feature aggregation*: the proposal features from the formed tube are aggregated, weighted by their semantic similarities with the current proposal candidate, enhancing the feature representations of the pedestrian in the current frame. Using the augmented features, the classifier tends to more confidently distinguish heavily occluded pedestrians from background. We implement this by a tube feature aggregation network (TFAN).

It is not straightforward to link heavily occluded pedestrians with non/less occluded ones, since their appearances are substantially different, otherwise most pedestrian detectors would deal well with occlusions. We resort to local spatial-temporal context to match pedestrians with different extents of occlusions using a new temporally discriminative embedding module (TDEM) and a part-based relation module (PRM). The TDEM module supervised by a discriminative loss learns an embedding for each proposal across frames, where pedestrian and background examples become readily separable in the embedding feature space. We therefore utilize these embedding features of proposals to search for their counterparts in consecutive frames and measure their semantic similarities as the weights to aggregate their features. When aggregating features from the tube, if the pedestrian proposal is heavily occluded, we favor the matched *reliable* pedestrians and assign them larger weights, rather than the backgrounds. However, the heavily occluded pedestrian may differ from the *reliable* ones due to missing observations. Accordingly, the PRM module is designed to focus more on the visible area of the current pedestrian candidate and assign the counterparts of similar visible parts with larger weights, so as to address the above discordance problem during feature aggregation.

The proposed TFAN strives to utilize local temporal context to enhance the feature representations of heavily occluded pedestrians by similar pedestrian samples in neighboring frames. Temporal clue has been widely exploited in video object detection. For instance, optical flow has been utilized to achieve feature calibration [30, 31, 38], while flow estimation may be noisy when an object is heavily occluded. Alternatively, detection boxes [33, 34, 37, 32] are associated to rerank classification scores as a post-processing step, yet these methods are not optimized end-to-end or require track-id annotations for training a tracker. By contrast, our approach integrates feature enhancement and pedestrian box association into a unified framework in an end-to-end fashion *without* the need of track-id annotations. Moreover, our approach is particularly designed for handling heavily occluded pedestrian detection.

In summary, our main contributions are three-fold: **(i)** We propose a tube feature aggregation network (TFAN), which essentially utilizes local temporal context to enhance the representations of heavily occluded pedestrians; **(ii)** We devise a temporally discriminative embedding module (TDEM) that links the tube reliably and assigns a robust and adaptive weight in aggregating tube features; **(iii)** We design a part-based relation module (PRM) which focuses on the visible pedestrian regions when aggregating features. Experiments on 3 benchmarks: Caltech [20], NightOwls [59] and KAIST [60] validate our approach is significantly effective for heavily occluded pedestrian detection.

2. Related Work

Pedestrian Detection. With the renaissance of convolutional neural networks, many deep learning based methods on pedestrian detection [27, 28, 10, 26, 24, 23, 19, 17, 25, 18, 48, 36] significantly outperform the hand-crafted feature based methods [55, 61, 21, 14]. Regardless of the promising performance on non-occluded pedestrians, most detectors yield limited accuracies on heavily occluded pedestrians. To alleviate the occlusion issue, recent methods are designed by exploiting attention mechanism [29, 9], feature transformation [11] and part-based detection [22, 19, 13]. Nevertheless, these works seldom take into account the temporal context, which may essentially help to compensate the missing information of heavily occluded pedestrians. To the best of our knowledge, TLL [23] is the only one recent work which also utilizes temporal cues for pedestrian detection. TLL simply applies an off-the-shelf LSTM [52] to the detection model. In contrast, our approach thoroughly investigates how to utilize local temporal context to enhance the representations of heavily occluded pedestrians.

Video Object Detection. Object detection in videos has been actively studied recently [50, 51, 38, 39, 40, 41, 42, 43, 30, 31, 38], exploring different ways to take advantage of temporal cues. Several works focus on utilizing optical flow to achieve feature calibration [30, 31, 38]. However, flow estimation may be inaccurate in the circumstance of fast motion. To tackle this problem, [44, 45, 49] propose to aggregate features at instance-level, which can better capture the objects with fast motion. Another direction is to associate proposal or detection boxes for tube classification and detection rescoring [34, 35, 33, 37, 32]. Nevertheless, these methods are not optimized end-to-end or require track-id annotations. In contrast, we present an end-to-end approach, integrating both proposal box association and feature augmentation into a unified framework *without* the need of track-id annotations. Since there may be mismatches in the linked tube, our approach performs a temporally discriminative embedding for each proposal across frames. When aggregating the tube features, only features from relevant counterparts are selected, so as to filter out irrelevant mismatches. Furthermore, our approach is dedicated to handling heavy occlusions in pedestrian detection, which has not been thoroughly investigated in the previous approaches.

3. Method

In this section, we first describe the baseline detector in § 3.1. Then, our proposed approach is presented in § 3.2. Finally, we introduce the implementation details in § 3.3.

3.1. Baseline Detector

For the baseline detector, we employ an off-the-shelf single-frame detector to process each frame individually in

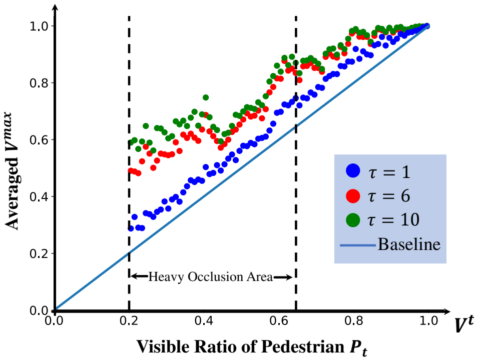

Figure 2. Visible ratio statistics of pedestrian examples on the Caltech dataset. For each pedestrian P_t in the current frame, we use the ground truth boxes to link a tube from $t-\tau$ to $t+\tau$ frames. The x-axis denotes the visible ratio of P_t. For each P_t, it has a V^{max} which is computed by the maximum visible ratio of those pedestrians in its corresponding tube. The y-axis denotes the average of the V^{max} for those P_t whose visible ratios are V^t. Pedestrians with visible ratios lower than 0.2 are not considered.

an input video. Specifically, we adopt vanilla Faster R-CNN [56] that is commonly used in pedestrian detection and ResNet-101 [57] of feature stride 16 as the base network.

3.2. Tube Feature Aggregation Network

In real-world scenarios, most pedestrians are actively moving and the heavily occluded ones are not always be occluded by other objects. To validate this, we conduct a quantitative analysis on the Caltech dataset as shown in Fig. 2. From the figure, we observe that most of pedestrians which are heavily occluded in the current frame become less occluded in nearby frames. Motivated by this observation, we aim to exploit local temporal context from neighboring frames to compensate the missing information of heavily occluded pedestrians.

3.2.1 Preliminary Model

Given a sequence of video frames $\{\mathbf{I}_i \in \mathbb{R}^{W\times H\times 3}\}_{i=t-\tau}^{t+\tau}$ where $\mathbf{I}_t$ is the current frame, we first apply the base network $\mathcal{N}_{\text{feat}}$ to each frame to produce feature maps $\mathbf{f}_i = \mathcal{N}_{\text{feat}}(\mathbf{I}_i)$, where $\mathbf{f}_i \in \mathbb{R}^{\frac{W}{16}\times\frac{H}{16}\times 256}$. Let us denote by $\mathcal{B}_i = \{\mathbf{b}_i^{k_i} \in \mathbb{R}^4\}_{k_i=1}^{M}$ the proposal boxes in frame $\mathbf{I}_i$ generated by the region proposal network [56] and $\mathcal{X}_i = \{\mathbf{x}_i^{k_i} \in \mathbb{R}^{7\times 7\times 256}\}_{k_i=1}^{M}$ the corresponding proposal features, where M $(= 300$ by default) is the total number of proposals per frame. $\mathbf{x}_i^{k_i}$ is obtained by $\mathbf{x}_i^{k_i} = \phi(\mathbf{f}_i, \mathbf{b}_i^{k_i})$, where ϕ is the RoI align operation [58]. In this paper, our goal is to enhance the proposal features $\mathcal{X}_t$ in the current frame, which is achieved by two steps: 1) *Tube linking*: starting from a pedestrian proposal $\mathbf{b}_t^{k_t}$, we iteratively search for its relevant counterparts in adjacent frames to form a proposal tube where we aim to include the *reliable* pedestrians in this tube; 2) *Feature aggregation*: the proposal features from the obtained tube are aggregated weighted by their semantic similarities with the current proposal candidate. Next, we introduce these two steps in detail.

Tube Linking. For simplicity, we only formulate the tube linking procedure from t to $t-\tau$, and the tube linking from t to $t+\tau$ is achieved in a similar way. Formally, let $\mathbf{b}_i^{k_i}$ denote the k_i-th proposal in frame $\mathbf{I}_i$. Starting from $\mathbf{b}_t^{k_t}$, we first look for its relevant counterparts in an adjacent spatial area in frame $\mathbf{I}_{t-1}$, and $\mathbf{b}_t^{k_t}$ is linked to the best matching counterpart $\mathbf{b}_{t-1}^{k_{t-1}}$ based on their semantic and spatial similarities. After $\mathbf{b}_{t-1}^{k_{t-1}}$ is found in frame $\mathbf{I}_{t-1}$, we then use it as the reference to search for the best matching counterpart $\mathbf{b}_{t-2}^{k_{t-2}}$ in frame $\mathbf{I}_{t-2}$. The linking procedure is iteratively performed until frame $\mathbf{I}_{t-\tau}$. Specifically, given the k_i-th proposal in frame $\mathbf{I}_i$, the best matching k_{i-1}-th proposal in frame $\mathbf{I}_{i-1}$ is found by:

$$k_{i-1} \;=\; \operatorname*{argmax}_{\hat{k}\in\mathcal{Q}_{k_{i-1}}} s(\mathbf{x}_i^{k_i}, \mathbf{x}_{i-1}^{\hat{k}}) + l(\mathbf{b}_i^{k_i}, \mathbf{b}_{i-1}^{\hat{k}}), \quad (1)$$

where $\mathcal{Q}_{k_{i-1}} = \{\hat{k} \mid IoU(\mathbf{b}_i^{k_i}, \mathbf{b}_{i-1}^{\hat{k}}) > \varepsilon\}$ is the set of indices of the proposals in frame $\mathbf{I}_{i-1}$ which are located in the adjacent spatial area of $\mathbf{b}_i^{k_i}$, and ε is a small constant that is set to 0.1 in experiments. $s(\cdot)$ and $l(\cdot)$ are the functions for measuring the semantic and spatial similarities between two proposals, respectively. Given two proposals $\mathbf{b}^1, \mathbf{b}^2$ and their corresponding proposal features $\mathbf{x}^1, \mathbf{x}^2$, the semantic similarity is measured by the cosine similarity between their proposal features:

$$s(\mathbf{x}^1, \mathbf{x}^2) = \frac{1}{|\mathcal{R}|}\sum_{p\in\mathcal{R}} \frac{\mathbf{x}^1(p)\cdot\mathbf{x}^2(p)}{|\mathbf{x}^1(p)||\mathbf{x}^2(p)|}, \quad (2)$$

where $\mathcal{R} = \{(x,y) \mid 1 \leqslant x \leqslant 7, 1 \leqslant y \leqslant 7\}$ is the set of spatial coordinates in the proposal features. The semantic similarity reflects the likelihood that two proposals belong to the same category. For the spatial similarity, we take into account both the scale and relative location information:

$$\begin{aligned} l(\mathbf{b}^1, \mathbf{b}^2) &= scale(\mathbf{b}^1, \mathbf{b}^2) + location(\mathbf{b}^1, \mathbf{b}^2), \\ scale(\mathbf{b}^1, \mathbf{b}^2) &= \min(\frac{w^1}{w^2}, \frac{w^2}{w^1}) \times \min(\frac{h^1}{h^2}, \frac{h^2}{h^1}), \\ location(\mathbf{b}^1, \mathbf{b}^2) &= exp(-\frac{\left\|(d_x^1, d_y^1) - (d_x^2, d_y^2)\right\|_2}{\sigma^2}), \end{aligned} \quad (3)$$

where w and h are the width and height of a proposal, respectively. d_x and d_y are predicted by the bounding box regression branch of Faster R-CNN, denoting the offset of the center of a proposal to its regression target. The term $scale(\cdot)$ is used to penalize a large scale change between two proposals in two consecutive frames, while the term $location(\cdot)$ is used to penalize a large mis-alignment between two proposals.

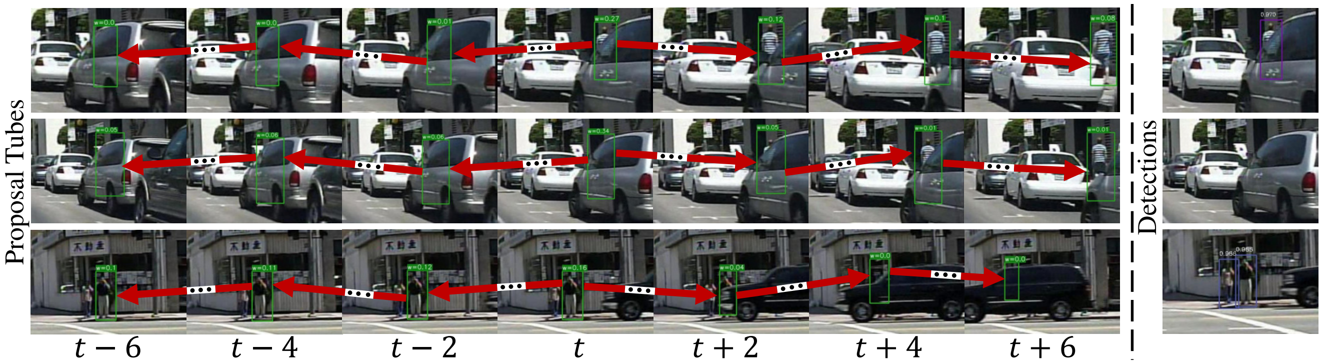

Figure 3. Visualization examples of the proposal tubes, adaptive weights and final detection results, where w in the figures denotes the adaptive weight. Three representative cases are provided, in which the current proposals are a heavily occluded pedestrian, a background region and a *reliable* pedestrian, respectively. For clear visualization, only one tube is shown in each row.

Finally, for the k_t-th proposal in the current frame, we obtain a proposal tube $\mathfrak{T}^{k_t}_{\mathbf{b},\tau} = \{\mathbf{b}^{k_{t-\tau}}_{t-\tau}, ..., \mathbf{b}^{k_t}_{t}, ..., \mathbf{b}^{k_{t+\tau}}_{t+\tau}\}$ and its corresponding tube features $\mathfrak{T}^{k_t}_{\mathbf{x},\tau} = \{\mathbf{x}^{k_{t-\tau}}_{t-\tau}, ..., \mathbf{x}^{k_t}_{t}, ..., \mathbf{x}^{k_{t+\tau}}_{t+\tau}\}$. Note that if $\mathbf{b}^{k_t}_t$ is a heavily occluded pedestrian and $\mathbf{b}^{k_{t-\tau}}_{t-\tau}$ is a non-occluded pedestrian, there is very likely a less occluded pedestrian in frame $\mathbf{I}_{t-\tau<i<t}$ due to temporal coherence. Therefore, in such a linking procedure, the less occluded pedestrian can serve as an intermediate step for building up the connection between $\mathbf{b}^{k_t}_t$ and $\mathbf{b}^{k_{t-\tau}}_{t-\tau}$, even if the direct semantic and spatial similarities between $\mathbf{b}^{k_t}_t$ and $\mathbf{b}^{k_{t-\tau}}_{t-\tau}$ may be not high.

Feature Aggregation. According to the analysis in Fig. 2, most heavily occluded pedestrians in the current frame may be related to some *reliable* (*i.e.*, non/less-occluded) counterparts in neighboring frames. By applying the iterative tube linking, we are able to connect the heavily occluded pedestrian in the current frame to the *reliable* ones in nearby frames. In view of these, we aggregate the proposal features from $\mathfrak{T}^{k_t}_{\mathbf{x},\tau}$ by a weighted summation, aiming at enhancing the current proposal features $\mathbf{x}^{k_t}_t$. Specifically, for proposal features $\mathbf{x}^{k_t}_t$, the enhanced features $\mathbf{x}^{k_t\prime}_t$ are computed by:

$$\mathbf{x}^{k_t\prime}_t = \sum_{i=t-\tau}^{t+\tau} w^{k_i}_i \mathbf{x}^{k_i}_i, \tag{4}$$

where $w^{k_i}_i$ is the adaptive weight and calculated as:

$$w^{k_i}_i = \frac{exp(\lambda \times s(\mathbf{x}^{k_t}_t, \mathbf{x}^{k_i}_i))}{\sum_{l=t-\tau}^{t+\tau} exp(\lambda \times s(\mathbf{x}^{k_t}_t, \mathbf{x}^{k_l}_l))}, \tag{5}$$

where λ is a scaling factor. Because the output value of $s(\cdot)$ is limited by cosine similarity which ranges from -1 to 1, λ is set to greater than 1 for enlarging the gap among examples. Considering there may be mismatches in the linked tube, we adopt the semantic similarity between $\mathbf{x}^{k_i}_i$ and $\mathbf{x}^{k_t}_t$ to determine the adaptive weight $w^{k_i}_i$, such that it can automatically select features from relevant counterparts and ignore some irrelevant or noisy ones once the tube drifts (see Fig. 3). Furthermore, we emphasize that the feature aggregation can augment not only the features of pedestrians but also those of backgrounds. If $\mathbf{b}^{k_t}_t$ is a background proposal, by tube linking, we are able to see more references around the nearby spatio-temporal areas, therefore facilitating the classifier to make a better decision and suppress false alarms.

3.2.2 Temporally Discriminative Embedding Module

In our preliminary model (§ 3.2.1), the tube linking and feature aggregation are mainly determined by the semantic similarity among proposal features. One issue is that pedestrian and background examples across frames may not be discriminative enough in the proposal feature space, as no explicit supervision is provided to enforce the proposal features of pedestrian and background examples to be separable. To address this, we learn a discriminative embedding $\mathbf{e}^{k_i}_i = \phi(\mathcal{N}_{\text{TDEM}}(\mathbf{f}_i), \mathbf{b}^{k_i}_i)$ for each proposal $\mathbf{b}^{k_i}_i$, where $\mathbf{e}^{k_i}_i \in \mathbb{R}^{7\times 7\times 256}$ and $\mathcal{N}_{\text{TDEM}}$ is the proposed temporally discriminative embedding module (TDEM) as shown in Fig. 4 (b). The $\mathcal{N}_{\text{TDEM}}$ is explicitly supervised by a discriminative loss L_{TDEM} , which enforces the pedestrian and background examples across frames to be more separable in the embedding feature space. Given the current frame $\mathbf{I}_t$ and a nearby frame $\mathbf{I}_i$, let us denote by $\mathcal{O} = \{\mathbf{e}^{k^\star_t}_t\}^U_{k^\star_t=1}$ the embedding features of the ground truth boxes in frame $\mathbf{I}_t$, where U is the number of ground truth boxes. For a ground truth box $\mathbf{b}^{k^\star_t}_t$ in the current frame, we denote $\mathbf{b}^{k^\star_i}_i$ as its corresponding ground truth box in frame $\mathbf{I}_i$, which is obtained by a greed scheme (as introduced in § 3.3). The L_{TDEM} is defined as:

$$L_{TDEM} = \frac{1}{|\mathcal{O}|} \sum_{\mathbf{e}^{k^\star_t}_t \in \mathcal{O}} \frac{1}{|\mathcal{Y}| \times |\mathcal{Z}|} \sum_{\mathbf{e}^n \in \mathcal{Y}, \mathbf{e}^p \in \mathcal{Z}} l_t(\mathbf{e}^n, \mathbf{e}^p, \mathbf{e}^{k^\star_t}_t), \tag{6}$$

where $\mathcal{Z}$ and $\mathcal{Y}$ are the sets of the embedding features of pedestrian and background proposals sampled around $\mathbf{b}^{k^\star_i}_i$,

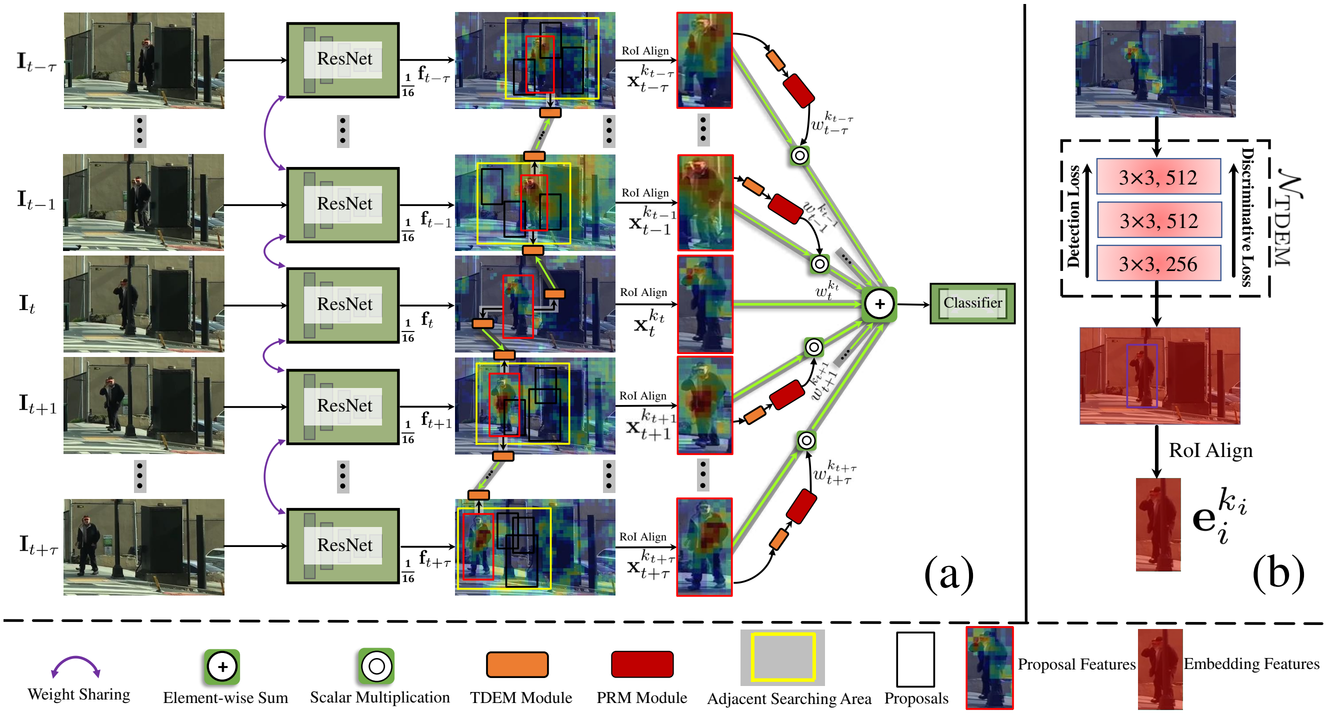

Figure 4. (a) Overall framework of the TFAN. Firstly, given an input video sequence, proposal tubes are formed based on the semantic and spatial similarities among proposals. Secondly, the proposal features from the obtained tube are aggregated according to the adaptive weights generated from the PRM module, enhancing the feature representations of the pedestrians in the current frame. Finally, the augmented proposal features are fed into two fully connected network layers for a better classification. (b) The proposed TDEM module, where the $\mathcal{N}_{TDEM}$ is learned by both detection loss and discriminative loss.

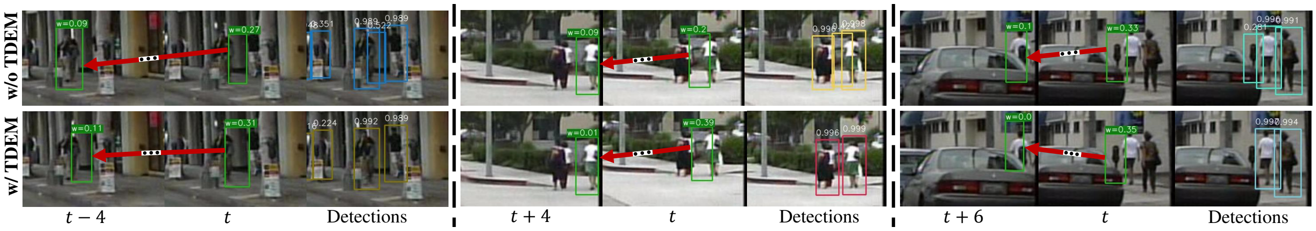

Figure 5. Qualitative examples of the TDEM module, where w in the figures denotes the adaptive weight. By applying a temporally discriminative embedding for each proposal, not only the drifting problem can be alleviated in linking tubes but also irrelevant mismatches are more effectively filtered out by the adaptive weights.

respectively, and $l_t(\cdot)$ is achieved by a triplet loss:

$$l_t(\mathbf{e}^n, \mathbf{e}^p, \mathbf{e}_t^{k_t^\star}) = \max(0, s(\mathbf{e}^n, \mathbf{e}_t^{k_t^\star}) - s(\mathbf{e}^p, \mathbf{e}_t^{k_t^\star}) + \alpha), \quad (7)$$

where the margin term α is set to 0.5 in experiments.

The discriminative embedding features learned from $\mathcal{N}_{\text{TDEM}}$ are then used for measuring the semantic similarity when linking tubes, which makes the TFAN be more likely to alleviate the drifting problem (as evidenced in Table 3). Moreover, such discriminative embedding features are further applied to each proposal in the formed tube for calculating the adaptive weights, so that it can more effectively absorb favorable features from relevant counterparts and filter out irrelevant mismatches (see Fig. 5). The adaptive weights can be also implicitly learned from $\mathcal{N}_{\text{TDEM}}$. With the discriminative embedding features, we rewrite Eq. 1 and Eq. 5 into:

$$k_{i-1} = \underset{\hat{k} \in \mathcal{Q}_{k_{i-1}}}{\operatorname{argmax}}\, s(\mathbf{e}_i^{k_i}, \mathbf{e}_{i-1}^{\hat{k}}) + l(\mathbf{b}_i^{k_i}, \mathbf{b}_{i-1}^{\hat{k}}), \quad (8)$$

$$w_i^{k_i} = \frac{exp(\lambda \times s(\mathbf{e}_t^{k_t}, \mathbf{e}_i^{k_i}))}{\sum_{l=t-\tau}^{t+\tau} exp(\lambda \times s(\mathbf{e}_t^{k_t}, \mathbf{e}_l^{k_l}))}. \quad (9)$$

3.2.3 Part-based Relation Module

Although a heavily occluded pedestrian $\mathbf{b}_t^{k_t}$ can be connected to a *reliable* pedestrian $\mathbf{b}_{t+\tau}^{k_{t+\tau}}$, the similarity $s(\mathbf{e}_t^{k_t}, \mathbf{e}_{t+\tau}^{k_{t+\tau}})$ may be small because the embedding features of the heavily occluded pedestrian are contaminated by background clutters. Accordingly, $\mathbf{x}_{t+\tau}^{k_{t+\tau}}$ will be overwhelmed by the proposal features of other examples when aggregating features. To better leverage those *reliable*

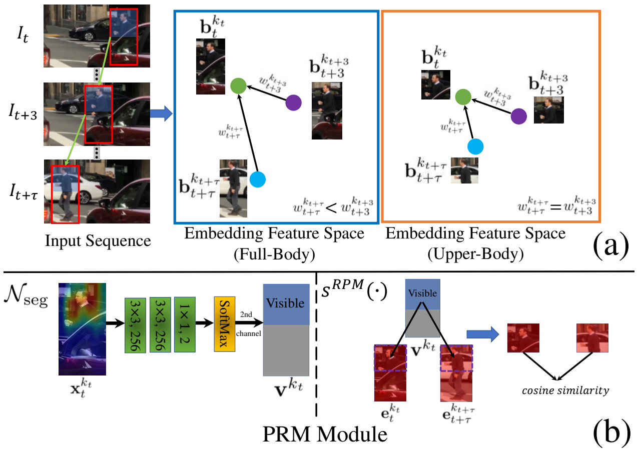
I_t
I_{t+3}
$I_{t+\tau}$
Input Sequence
$\mathbf{b}_t^{k_t}$
$\mathbf{b}_{t+3}^{k_{t+3}}$
$w_{t+3}^{k_{t+3}}$
$w_{t+\tau}^{k_{t+\tau}}$
$\mathbf{b}_{t+\tau}^{k_{t+\tau}}$
$w_{t+\tau}^{k_{t+\tau}} < w_{t+3}^{k_{t+3}}$
Embedding Feature Space (Full-Body)
$\mathbf{b}_t^{k_t}$
$\mathbf{b}_{t+3}^{k_{t+3}}$
$w_{t+3}^{k_{t+3}}$
$w_{t+\tau}^{k_{t+\tau}}$
$\mathbf{b}_{t+\tau}^{k_{t+\tau}}$
$w_{t+\tau}^{k_{t+\tau}} = w_{t+3}^{k_{t+3}}$
Embedding Feature Space (Upper-Body)
(a)
$\mathcal{N}_{\text{seg}}$
$\mathbf{x}_t^{k_t}$
3×3, 256
3×3, 256
1× 1, 2
SoftMax
2nd
channel
Visible
$\mathbf{v}^{k_t}$
$s^{RPM}(\cdot)$
Visible
$\mathbf{v}^{k_t}$
$\mathbf{e}_t^{k_t}$
$\mathbf{e}_{t+\tau}^{k_{t+\tau}}$
cosine similarity
PRM Module
(b)

Figure 6. (a) Illustration of the motivation of the PRM module. (b) The proposed PRM module.

pedestrians, we design a part-based relation module (PRM) as shown in Fig. 6 (b). For a current pedestrian candidate, the PRM module will favor its counterparts with similar visible parts and assign them large adaptive weights in aggregating features. For the example in Fig. 6 (a), we want to use the embedding features of upper-body to measure the semantic similarity between $\mathbf{b}_t^{k_t}$ and $\mathbf{b}_{t+\tau}^{k_{t+\tau}}$, since both their upper parts are visible. To this end, given a pair of $\mathbf{b}_t^{k_t}$ and $\mathbf{b}_i^{k_i}$, the PRM module first applies a segmentation subnetwork $\mathcal{N}_{\text{seg}}$ to $\mathbf{x}_t^{k_t}$ to predict the visible mask $\mathbf{v}^{k_t} = \mathcal{N}_{\text{seg}}(\mathbf{x}_t^{k_t})$ for the current pedestrian candidate, where $\mathbf{v}^{k_t} \in [0,1]^{7\times 7\times 1}$. Next, the adaptive weight $w_i^{k_i}$ is computed using an improved semantic similarity function $s^{PRM}(\cdot)$, which is defined in terms of $\mathbf{v}^{k_t}$:

$$s^{PRM}(\mathbf{e}_t^{k_t}, \mathbf{e}_i^{k_i}) = \frac{1}{|\mathcal{V}|} \sum_{p \in \mathcal{V}} \frac{\mathbf{e}_t^{k_t}(p) \cdot \mathbf{e}_i^{k_i}(p)}{|\mathbf{e}_t^{k_t}(p)||\mathbf{e}_i^{k_i}(p)|}, \tag{10}$$

where $\mathcal{V} = \{p \mid \mathbf{v}^{k_t}(p) \geqslant \min\{0.5, \gamma\}\}$ and γ is a threshold which adaptively determined by $\mathbf{v}^{k_t}$. For background, the values in $\mathbf{v}^{k_t}$ tend to be zero. In order to retain enough pixels for computing the semantic similarity for background proposals, γ is set to a value such that at least 20% pixels in embedding features are retained. The percentile 20% is chosen according to the definition of heavy occlusion in existing pedestrian dataset: a pedestrian is considered to be heavily occluded if only $20\% - 65\%$ of its body is visible.

3.2.4 Discussion

The overall architecture of the TFAN is shown in Fig. 4 (a). The TFAN is designed to exploit the local spatial-temporal context of heavily occluded pedestrians to enhance their representations in the current frame. Different from person tracking, the TFAN does not necessarily require the proposals in the linked tube $\mathfrak{T}_{\mathbf{b},\tau}^{k_t}$ with the same pedestrian identity, and instances from different persons may also contribute to augment the $\mathbf{x}_t^{k_t}$ as long as they have distinguishable feature representations. Moreover, our model also enjoys the enhanced discriminability of background features. For pedestrian detection especially in night time, some ambiguous negative examples, *e.g.*, trees and poles, are often misclassified with a high confidence score by the single frame detector. In our approach, we are able to utilize more samples around nearby spatio-temporal areas, so that these hard negative examples are confidently suppressed by the classifier (as shown in *Supplementary Material*).

3.3. Implementation

Training. The proposed TFAN is fully differentiable and can be trained end-to-end. Similar to [30], we select 3 frames $\mathbf{I}_{bef}, \mathbf{I}_t, \mathbf{I}_{aft}$ for training due to limited memory, where $\mathbf{I}_{bef}$ and $\mathbf{I}_{aft}$ are randomly sampled from $\{\mathbf{I}_i\}_{i=t-\tau}^{t-1}$ and $\{\mathbf{I}_i\}_{i=t+1}^{t+\tau}$, respectively. The overall loss function of the TFAN is defined as:

$$L = L_{det} + L_{seg} + L_{TDEM}, \tag{11}$$

where L_{det} is the detection loss for Faster R-CNN as in [56], L_{seg} is the segmentation loss for $\mathcal{N}_{\text{seg}}$ and L_{TDEM} is the discriminative loss for $\mathcal{N}_{\text{TDEM}}$. Cross-entropy loss is used for L_{seg}. Since pixel-level annotations for visible pedestrian areas are not available in existing pedestrian detection datasets, we use the visible bounding boxes as a weak supervision for $\mathcal{N}_{\text{seg}}$ as in [29]. For L_{TDEM}, we need to find those ground truth boxes in frames $\mathbf{I}_{bef}$ and $\mathbf{I}_{aft}$ which correspond to the ground truth box $\mathbf{b}_t^{k_t^\star}$ in frame $\mathbf{I}_t$. Since the track-id annotations are unavailable in some pedestrian detection datasets, we adopt a greedy scheme to obtain them. Specifically, starting from $\mathbf{b}_t^{k_t^\star}$, we iteratively find the corresponding one in next frame using IoU as a matching score until $\mathbf{I}_{bef}$ or $\mathbf{I}_{aft}$ are reached.

Inference. Given the input video frames $\{\mathbf{I}_i\}_{i=t-\tau}^{t+\tau}$ ($\tau = 6$ by default), our approach outputs the detection boxes in frame $\mathbf{I}_t$. In our implementation, we decouple the branches of classification and bounding box regression. For classification, we use the enhanced features $\mathbf{x}_t^{k_t\prime}$. For bounding box regression, the original $\mathbf{x}_t^{k_t}$ are used.

4. Experiments

4.1. Datasets and Experiment Settings

Dataset. In order to exploit temporal cue in our approach, we conduct experiments on three large-scale pedestrian detection datasets: Caltech [20], NightOwls [59] and KAIST [60], where the video sequences are publicly available. On the Caltech dataset, the results are reported on three subsets: Reasonable (R), Heavy Occlusion (HO) and Reasonable+Heavy Occlusion (R+HO), where the visible ratios of pedestrians are in the range of $[0.65, 1]$, $[0.2, 0.65]$

Method	R+HO	HO	R
Baseline	16.5	43.1	8.6
Baseline+FGFA[30]	15.7	38.9	8.2
SELSA[44]	14.9	39.6	7.5
TFAN-preliminary	14.2	37.6	7.2
TFAN+TDEM (w/o L_{TDEM})	14.1	37.5	7.0
TFAN+TDEM (w/o *spa*)	13.0	33.5	6.9
TFAN+TDEM+PRM (w/o *spa*)	13.0	33.2	6.8
TFAN+TDEM	12.9	32.7	6.8
TFAN+TDEM+PRM	**12.4**	**30.9**	**6.7**

Table 1. Ablation study of each proposed module on the Caltech dataset. w/o *spa* denotes that the TFAN achieves the tube linking without considering the spatial similarity.

NightOwls			
Subset	Baseline	Ours	Δ
Occluded	46.5	**42.1**	+4.4
Reasonable+Occluded	20.8	**18.5**	+2.3
Reasonable	16.3	**14.3**	+2.0
KAIST			
Subset	Baseline	Ours	Δ
Heavy Occlusion	76.6	**71.3**	+5.3
Partial Occlusion	55.4	**49.0**	+6.4
Reasonable	35.9	**34.6**	+1.3

Table 2. Performance comparison with the baseline detector on the NightOwls validation set and the KAIST testing set, respectively. Ours indicates the TFAN+TDEM.

and $[0.2, 1]$, respectively. NightOwls is a newly released dataset, in which all the images are captured in night time. As the NightOwls dataset only provides a binary occlusion flag in annotations, we report the results on the Reasonable, Occluded and Reasonable+Occluded subsets. KAIST mainly focuses on multispectral pedestrian detection, in which half of the images are also collected in night time. Following the common protocols, we experiment on three subsets: Reasonable, Partial Occlusion and Heavy Occlusion whose pedestrian examples have visible ratios in the range of $[0.5, 1]$, $[0.5, 1)$ and $[0, 0.5]$, respectively. On the KAIST dataset, the stat-of-the-art methods mainly work on the fusion of thermal images and RGB images, which is not our focus in this paper. Therefore, we only use RGB images and compare our approach with the baseline detector on the KAIST dataset. The original annotations of these three datasets are used for experiments.

Experiment Settings. We adopt the standard evaluation metric in pedestrian detection: MR^{-2} (lower is better). The TFAN is trained with 3 epochs using SGD optimizer, and the initial learning rate is set to 0.0005 and decreased by a factor of 10 after 2 epochs. σ and λ are respectively set to 0.5 and 5 by default.

4.2. Ablation Studies

Comparison with Baselines. As shown in Table 1, we compare three variants of our approach: TFAN-preliminary (§ 3.2.1), TFAN+TDEM (§ 3.2.2) and TFAN+TDEM+PRM (§ 3.2.3) as well as the baseline detector. To better compare with video object detection methods, we also list the results of FGFA [30] and SELSA [44]. Compared with the baseline detector, our proposed approach boosts the detection performance on HO subset by a remarkably large margin of 12.2 points. Besides, we observe that our approach also improves the detection performance on the *reliable* pedestrians. On the NightOwls and KAIST datasets, we use the TFAN+TDEM for experiments due to the lack of visible region annotations. Table 2 shows that our approach is effective for heavily occluded pedestrians even in the night scenario, showing very well generalization ability of the proposed method. Qualitative detection performance on the heavily occluded pedestrians can be found in *Supplementary Material*. In the following ablation studies, experiments are analyzed on the Caltech dataset.

KNN	K= 1	K= 3	K= 5	K= 7
proposal feature space	68.3	70.2	69.8	66.6
embedding feature space	**72.8**	**77.0**	**80.5**	**79.8**

Table 3. Classification accuracy of pedestrian proposals by KNN using the proposal features and embedding features, respectively.

Effectiveness of the TDEM Module. Table 1 shows that the TDEM module (§ 3.2.2) mainly benefits from the discriminative loss rather than the additional network layers. To further quantitatively analyze the proposed TDEM module in depth, we utilize K-nearest neighbors algorithm (KNN) to classify the pedestrian examples in both the proposal feature space and embedding feature space. Specifically, given an input image with 300 proposals, a proposal is classified by a plurality vote of its neighbors in the feature space. Euclidean distance is used for measuring the distance between two proposals in the feature space. In Table 3, we report the classification accuracy with different K on the Caltech testing set, where the accuracy is the percentage of correctly classified pedestrian proposals over the total pedestrian proposals. It is clear to see that pedestrians and backgrounds are more separable in the proposed embedding feature space, which can benefit both the tube linking and feature aggregation (see Fig. 5).

Effectiveness of the PRM Module. To study the effect of the PRM module (§ 3.2.3), we visualize the predicted visible masks and cosine similarity maps which are used to measure the semantic similarity. As shown in Fig. 7, when using the embedding features of full-body to measure the semantic similarity (Eq. 2), the heavily occluded pedestrians are relatively less similar to the *reliable* ones. By focusing on the visible parts of current pedestrian candidates (Eq. 10), the PRM module is more likely to recall those *reliable* pedestrians for supporting the heavily occluded ones in the current frame. Besides, we can see from Table 1 that the

Subset	Ours	Baseline	SDS-RCNN [7]	RPN+BF[2]	A-FRCCN[6]	FRCCN[53]	Checkerboards[21]
Reasonable	**16.5**	19.7	**17.8**	23.3	18.8	20.0	39.7

Table 4. Performance comparison with the state-of-the-art methods on the NightOwls testing subset. Ours indicates the TFAN+TDEM.

τ	3	4	5	6	7	8	9	10
R+HO	13.0	12.8	12.5	**12.4**	**12.4**	**12.4**	**12.4**	**12.4**
HO	32.9	32.2	31.5	**30.9**	31.1	31.2	31.5	31.6
R	7.1	6.9	6.9	6.7	6.7	6.6	**6.5**	6.6

Table 5. Ablation study of the TFAN with different tube lengths.

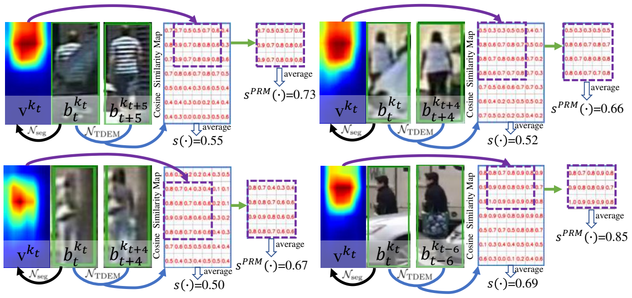

Figure 7. Qualitative examples of the PRM module.

PRM module is more effective when taking into consideration the spatial similarity in linking tubes. The main reason is that the similarity measurement is conducted using the visible parts of current pedestrian candidates. Therefore, a spatially aligned tube shall be more beneficial for this kind of measurement.

Adaptive Weights. To assess the effectiveness of the adaptive weights, we experiment our detector with the average weights, *i.e.*, $w_i^{k_i} = \frac{1}{\tau}$. The TFAN with average weights achieves $14.9/35.0/8.4$ MR^{-2} on the R+HO, HO and R subsets, respectively, where the TFAN with the adaptive weights obtains $12.4/30.9/6.7$ MR^{-2}. The performance degradation is that the average weights may not adaptively filter out irrelevant features during feature aggregation.

Tube Length. We experiment our approach with different tube lengths from $\tau = 3$ to 10. As shown in Table 5, performance tends to be stable when $\tau \geqslant 5$, indicating that our method does not require a long tube and 11 frames are enough to support the detection in the current frame.

Hyper-parameters. There are several hyper-parameters in the proposed method, *e.g.*, σ, λ, γ. The results of our approach with different hyper-parameters can be found in *Supplementary Material*, which shows our approach is not sensitive to these hyper-parameters.

4.3. Comparison with State-of-the-Art

Caltech Dataset. We list the state-of-art methods which use no extra data in Table 6. Our approach achieves notable performance improvements on the R+HO and HO subsets, respectively, outperforming the second best results by 1.5 and 6.4 points. It shows our detector is specialized to detect heavily occluded pedestrians.

Method	Occ	R+HO	HO	R
CompACT-Deep [1]		24.6	65.8	11.7
RPN+BF [2]		24.0	74.4	9.6
DeepParts [3]	✓	22.8	60.4	11.9
SAF-RCNN [4]		21.9	64.4	9.7
MS-CNN [5]		21.5	59.9	10.0
A-FRCNN [6]		20.0	57.6	9.2
SDS-RCNN [7]		19.7	58.5	7.4
F-DNN [8]		19.3	55.1	8.6
ATT-part [9]	✓	18.2	45.2	10.3
AR-Ped [10]		16.1	48.8	**6.5**
Bi-Box [12]	✓	16.1	44.4	7.6
DSSD+Grid [19]	✓	-	42.42	10.9
GDFL [17]	✓	15.6	43.2	7.8
FRCN+A+DT [11]	✓	15.2	**37.9**	8.0
MGAN [29]	✓	**13.9**	38.3	**6.8**
TFAN+TDEM+PRM	✓	**12.4**	**31.5**	**6.5**

Table 6. Performance comparison with the state-of-the-art methods on the Caltech dataset. The Occ column indicates whether an approach is devised for handling occlusions. The top two scores are highlighted in **red** and **blue**, respectively.

NightOwls Dataset. We compare the state-of-the-art methods on the NightOwls testing subset, where only evaluation on the Reasonable subset is publicly available. As shown in Table 4, the proposed method outperforms the second best result by 1.3 points on the Reasonable subset, validating that our approach can be well generalized to night time scenario.

5. Conclusion

This work presents a novel model, the TFAN, aiming at exploiting local spatial and temporal context of a heavily occluded pedestrian to enhance its feature representations. The TFAN is carried out with two main steps: tube linking and feature aggregation, which are designed to search for relevant counterparts temporally in the video and exploit them to enhance the feature representations of current pedestrian candidates. Furthermore, the TFAN together with the TDEM and PRM modules is capable of handling drifting and severe occlusion problems. Extensive experiments validate the effectiveness and superiority of our method.

Acknowledgment. This work is supported in part by the gift grant from Horizon Robotics and start-up funds from State University of New York at Buffalo.

References

[1] Z. Cai, M. Saberian, and N. Vasconcelos. Learning complexity-aware cascades for deep pedestrian detection. In *ICCV*, 2015. 8

[2] L. Zhang, L. Lin, X. Liang, and K. He. Is faster r-cnn doing well for pedestrian detection? In *ECCV*, 2016. 8

[3] Y. Tian, P. Luo, X. Wang, and X. Tang. Deep learning strong parts for pedestrian detection. In *ICCV*, 2015. 8

[4] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan. Scale-aware fast r-cnn for pedestrian detection. *IEEE TMM*, 2017. 8

[5] Z. Cai, Q. Fan, R. S. Feris, and N. Vasconcelos. A unified multi-scale deep convolutional neural network for fast object detection. In *ECCV*, 2016. 8

[6] S. Zhang, R. Benenson, and B. Schiele. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*, 2017. 8

[7] G. Brazil, X. Yin, and X. Liu. Illuminating pedestrians via simultaneous detection and segmentation. In *ICCV*, 2017. 8

[8] X. Du, M. El-Khamy, J. Lee, and L. Davis. Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection. In *WACV*, 2017. 8

[9] S. Zhang, J. Yang, and B. Schiele. Occluded pedestrian detection through guided attention in cnns. In *CVPR*, 2018. 1, 2, 8

[10] G. Brazil, and X. Liu. Pedestrian detection with autoregressive network phases. In *CVPR*, 2019. 1, 2, 8

[11] C. Zhou, M. Yang, and J. Yuan. Discriminative feature transformation for occluded pedestrian detection. In *ICCV*, 2019. 1, 2, 8

[12] C. Zhou, and J. Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *ECCV*, 2018. 8

[13] C. Zhou, and J. Yuan. Multi-label learning of part detectors for heavily occluded pedestrian detection. In *ICCV*, 2017. 1, 2

[14] C. Zhou, and J. Yuan. Learning to integrate occlusion-specific detectors for heavily occluded pedestrian detection. In *ACCV*, 2016. 2

[15] C. Zhou, and J. Yuan. Multi-label learning of part detectors for occluded pedestrian detection. *PR*, 2019. 1

[16] C. Zhou, and J. Yuan. Occlusion Pattern Discovery for Partially Occluded Object Detection. *IEEE TCSVT*, 2020. 1

[17] C. Lin, J. Lu, G. Wang, and J. Zhou. Graininess-aware deep feature learning for pedestrian detection. In *ECCV*, 2018. 1, 2, 8

[18] J. Mao, T. Xiao, Y. Jiang, and Z. Cao. What can help pedestrian detection? In *CVPR*, 2017. 2

[19] J. Noh, S. Lee, B. Kim, and G. Kim. Improving occlusion and hard negative handling for single-stage pedestrian detectors. In *CVPR*, 2018. 1, 2, 8

[20] P. Dollár, C. Wojek, B. Schiele, and P. Perona. Pedestrian detection: An evaluation of the state of the art. *IEEE TPAMI*, 2012. 2, 6

[21] S. Zhang, R. Benenson, and B. Schiele. Filtered channel features for pedestrian detection. In *CVPR*, 2015. 2, 8

[22] S. Zhang, L. Wen, X. Bian, Z. Lei, and S. Z. Li. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In *ECCV*, 2018. 1, 2

[23] T. Song, L. Sun, D. Xie, H. Sun, and S. Pu. Small-scale pedestrian detection based on somatic topology localization and temporal feature aggregation. In *ECCV*, 2018. 1, 2

[24] W. Liu, S. Liao, W. Hu, X. Liang, and X. Chen. Learning efficient single-stage pedestrian detectors by asymptotic localization fitting. In *ECCV*, 2018. 1, 2

[25] W. Ouyang, and X. Wang. Joint deep learning for pedestrian detection. In *ICCV*, 2013. 2

[26] X. Wang, T. Xiao, Y. Jiang, S. Shao, J. Sun, and C. Shen. Repulsion loss: Detecting pedestrian in a crowd. In *CVPR*, 2018. 1, 2

[27] S. Liu, D. Huang, and Y. Wang. Adaptive nms: refining pedestrian detection in a crowd. In *CVPR*, 2019. 1, 2

[28] W. Liu, S. Liao, W. Ren, W. Hu, and Y. Yu. High-level semantic feature detection: a new perspective for pedestrian detection. In *CVPR*, 2019. 1, 2

[29] Y. Pang, J. Xie, M. H. Khan, R. M. Anwer, F. S. Khan, and L. Shao. Learning to mask visible regions for occluded pedestrian detection. In *ICCV*, 2019. 1, 2, 6, 8

[30] X. Zhu, Y. Wang, J. Dai, L. Yuan, and Y. Wei. Flow-guided feature aggregation for video object detection. In *ICCV*, 2017. 2, 6, 7

[31] X. Zhu, Y. Xiong, J. Dai, L. Yuan, and Y. Wei. Deep feature flow for video recognition. In *CVPR*, 2017. 2

[32] C. Feichtenhofer, A. Pinz, and A. Zisserman. Detect to track and track to detect. In *ICCV*, 2017. 2

[33] K. Kang, H. Li, J. Yan, X. Zeng, B. Yang, T. Xiao, C. Zhang, Z. Wang, R. Wang, X. Wang, and W. Ouyang. T-cnn: tubelets with convolutional neural networks for object detection from videos. *IEEE TCSVT*, 2017. 2

[34] K. Kang, W. Ouyang, H. Li, and X. Wang. Object detection from video tubelets with convolutional neural networks. In *CVPR*, 2016. 2

[35] K. Kang, W. Ouyang, H. Li, and X. Wang. K. Kang, H. Li, T. Xiao, W. Ouyang, J. Yan, X. Liu, and X. Wang. Object detection in videos with tubelet proposal networks. In *CVPR*, 2017. 2

[36] J. Cao, Y. Pang, S. Zhao, and X. Li. High-level semantic networks for multi-Scale object detection. *IEEE TCSVT*, 2019. 2

[37] W. Han, P. Khorrami, T. Paine, P. Ramachandran, M. Babaeizadeh, H. Shi, J. Li, S. Yan, and T. Huang. Seq-NMS for video object detection. *arXiv:1602.08465*, 2016. 2

[38] S. Wang, Y. Zhou, J. Yan, and Z. Deng. Fully motion-aware network for video object detection. In *ECCV*, 2018. 2

[39] F. Xiao, and Y. J. Lee. Video object detection with an aligned spatial-temporal memory. In *ECCV*, 2018. 2

[40] X. Zhu, J. Dai, L. Yuan, and Y. Wei. Towards high performance video object detection. In *CVPR*, 2018. 2

[41] X. Wang, and A. Gupta. Videos as space-time region graphs. In *ECCV*, 2018. 2

[42] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *CVPR*, 2018. 2

[43] K. Chen, J. Wang, S. Yang, X. Zhang, Y. Xiong, and C. C. Loy, and D. Lin. Optimizing video object detection via a scale-time lattice. In *CVPR*, 2018. 2

[44] H. Wu, Y. Chen, N. Wang, and Z. Zhang. Sequence level semantics aggregation for video object detection. In *ICCV*, 2019. 2, 7

[45] M. Shvets, W. Liu, and A. Berg. Leveraging long-range temporal relationships between proposals for video object detection. In *ICCV*, 2019. 2

[46] J. Cao, Y. Pang, J. Han, B. Gao, and X. Li. Taking a look at small-scale pedestrians and occluded pedestrians. *IEEE TIP*, 2019. 1

[47] J. Cao, Y. Pang, and X. Li. Pedestrian detection inspired by appearance constancy and shape symmetry. In *CVPR*, 2016. 1

[48] J. Cao, Y. Pang, and X. Li. Learning multilayer channel features for pedestrian detection. *IEEE TIP*, 2016. 2

[49] J. Deng, Y. Pan, Ting. Yao, W. Zhou, H. Li, and T. Mei. Relation distillation networks for video object detection. In *ICCV*, 2019. 2

[50] C. Guo, B. Fan, J. Gu1, Q. Zhang, S. Xiang, V. Prinet, and C. Pan. Progressive sparse local attention for video object detection. In *ICCV*, 2019. 2

[51] H. Deng, Y. Hua, T. Song, Z. Zhang, Z. Xue1, R. Ma, N. Robertson, and H. Guan. Object guided external memory network for video object detection. In *ICCV*, 2019. 2

[52] X. Shi, Z. Chen, H. Wang, D. Yeung,W. Wong, and W. Woo. Convolutional lstm network: a machine learning approach for precipitation nowcasting. In *NeurIPS*, 2015. 2

[53] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 8

[54] J. Xie, Y. Pang, H. Cholakkal, R. M. Anwer, F. S. Khan, and L. Shao. Psc-net: learning part spatial co-occurrence for occluded pedestrian detection. *arXiv:2001.09252*, 2020. 1

[55] P. Dollár, R. Appel, S. Belongie, and P. Perona. Fast feature pyramids for object detection. *IEEE TPAMI*, 2014. 2

[56] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3, 6

[57] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[58] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017. 3

[59] 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, 6

[60] S. Hwang, J. Park, N. Kim, Y. Choi, and I. Kweon. Multispectral pedestrian detection: benchmark dataset and baseline. In *CVPR*, 2015. 2, 6

[61] P. Dollár, Z. Tu, P. Perona, and S. Belongie. Integral channel features. In *BMVC*, 2009. 2