Information Fusion 46 (2019) 206–217

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Pedestrian detection with unsupervised multispectral feature learning using deep neural networks

Yanpeng Cao[a,b], Dayan Guan[b], Weilin Huang[c], Jiangxin Yang[a,b], Yanlong Cao[a,b], Yu Qiao[*,d]

[a] State Key Laboratory of Fluid Power and Mechatronic Systems, School of Mechanical Engineering, Zhejiang University, Hangzhou, China
[b] Key Laboratory of Advanced Manufacturing Technology of Zhejiang Province, School of Mechanical Engineering, Zhejiang University, Hangzhou, China
[c] Malong Technologies Co., Ltd., Shenzhen, China
[d] Key Lab of Comp. Vis and Pat. Rec. of Guangdong Province, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

ARTICLE INFO

Keywords:
Multispectral pedestrian detection
Deep neural networks
Auto-annotation
Semantic feature fusion
Unsupervised learning

ABSTRACT

Multispectral pedestrian detection is an important functionality in various computer vision applications such as robot sensing, security surveillance, and autonomous driving. In this paper, our motivation is to automatically adapt a generic pedestrian detector trained in a visible source domain to a new multispectral target domain without any manual annotation efforts. For this purpose, we present an auto-annotation framework to iteratively label pedestrian instances in visible and thermal channels by leveraging the complementary information of multispectral data. A distinct target is temporally tracked through image sequences to generate more confident labels. The predicted pedestrians in two individual channels are merged through a label fusion scheme to generate multispectral pedestrian annotations. The obtained annotations are then fed to a two-stream region proposal network (TS-RPN) to learn the multispectral features on both visible and thermal images for robust pedestrian detection. Experimental results on KAIST multispectral dataset show that our proposed unsupervised approach using auto-annotated training data can achieve performance comparable to state-of-the-art deep neural networks (DNNs) based pedestrian detectors trained using manual labels.

1. Introduction

Pedestrian detection has attracted much attention within computer vision community in recent years [1–5]. Given images captured under varying conditions of illumination, viewpoints, camera resolutions and backgrounds, pedestrian detection solution is required to generate bounding boxes to predict accurate positions of individual pedestrian instances. It provides an important functionality for various human-centric applications, such as autonomous vehicles, urban monitoring, UAV surveillance and intelligent robots.

Although significant improvements have been accomplished over the past few years, it still remains a difficult problem to train a pedestrian detector that works reliably in various real-world surveillance situations. The major challenge is two-fold. First, the performance of a generic pedestrian detector may drop significantly when it is applied to new scenes, due to the mismatch between the source and target scenes [6]. For example, as shown in Fig. 1(a) and (b), a state-of-the-art region proposal network (RPN) detector [7] well-trained in the Caltech pedestrian dataset [8] cannot produce satisfactory detection results when applied to visible images from the KAIST pedestrian benchmark, even under a similar lighting condition. Second, most pedestrian detectors are trained on RGB images captured by visible cameras which are sensitive to changes of illumination, weather and occlusions. A detector built on such images is very likely to be stuck with images captured during nighttime. As shown in Fig. 1(c), many obvious pedestrians cannot be properly localized using a RPN detector. The situation is even worse in a dark condition where pedestrians may become unrecognizable in visible spectrum, as shown in Fig. 1(d). Recently, pedestrian detection under extreme conditions becomes increasingly important for around-the-clock robotic applications, such as autonomous vehicle and surveillance system [9–11]. This imposes the applications of multimodal data sources for pedestrian detection [12,13], which is able to provide multi-cue information about objects of interest, leading to more robust and reliable detection.

To achieve a high detection accuracy in target scenes, it is crucial to train a pedestrian detector on target data with multiple-cue information incorporated. Consequently, this often requires large-scale annotations on the target data, which is costly and unscalable, since the target

* Corresponding author.
E-mail address: yu.qiao@siat.ac.cn (Y. Qiao).

https://doi.org/10.1016/j.inffus.2018.06.005
Received 30 September 2017; Received in revised form 11 April 2018; Accepted 19 June 2018
Available online 20 June 2018

Fig. 1. Detection results of a generic RPN detector well-trained using images from Caltech dataset. (a) Results on an image from Caltech dataset; (b) Results on an image from the KAIST dataset captured during daytime; (c) Results on an image from the KAIST dataset captured during nighttime in good lighting condition; (d) Results on an image from the KAIST dataset captured during nighttime in dark condition. Note green bounding boxes show true positives, red bounding boxes show false positives, and red bounding boxes in dashed line show false negatives. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

domain might change frequently. Furthermore, reliable and consistent manual annotations over various image modalities are difficult, since visual information in some models may be relatively weak, such as visible nighttime images. In this work, we propose a novel framework for unsupervised learning of a DNN-based pedestrian detector, by leveraging the complementary information of multispectral data. We present a number of technical developments that allow our detector to be readily adapted to new target scenes, without any manual annotation. We show in experiments that the proposed unsupervised approach achieves performance comparable to methods based on supervised DNN-based detector. The contributions of our work are three-fold.

First, we demonstrate the usefulness of multispectral data for the task of pedestrian detection. Complementary information captured by multimodal sensors (visible and infrared cameras) can not only be used to improve detection accuracy but also to automatically generate training samples. To the best of our knowledge, this is the first attempt to explore visible and infrared images for learning multispectral pedestrian detector in an unsupervised manner.

Second, we propose a novel approach able to iteratively label pedestrian instances from visible and thermal channels. It takes ~ 30 hours to complete the auto-annotation of 50k multispectral image pairs while the manual labeling process of such data requires more than 80 hours. These auto-annotation results can be used to train a detector, making our approach readily applicable to new target scenes without extra manual labeling efforts.

Third, we present a unified framework which combines the auto-annotation method with a TS-RPN detector to achieve unsupervised learning of multispectral features for robust pedestrian detection. On the KAIST multispectral dataset which captures various urban scenes at both daytime and nighttime, our proposed unsupervised approach using auto-annotated training data alone can achieve performance comparable to state-of-the-art DNNs-based multispectral pedestrian detectors trained using manual labels (36.42% vs. 36.99% on miss rates).

2. Related work

Researches closely related to our work include supervised pedestrian detection, multimodal data fusion and unsupervised domain adaptation. We present a review of the most recent works on these topics below.

Supervised pedestrian detection: a large variety of methods have been proposed for improving pedestrian detection systems to achieve faster speed and higher accuracy. Papageorgiou et al. [14] proposed the first sliding window detectors, applying support vector machines (SVM) to multi-scale Haar wavelets. Viola and Jones [15] introduced integral images for fast feature computation and a cascade structure for efficient detection, utilizing AdaBoost for automatic feature selection. Piotr et al. [16] built the Integrate Channel Features (ICF) detector with channel feature pyramids and boosted classifiers. The feature representations of ICF have been improved through multiple techniques, including aggregated channel features (ACF) [8], locally decorrelated channel features (LDCF) [17], Checkerboards [18] etc. Recently DNNs-based approaches have been widely applied to improve the performance of pedestrian detection [19–22]. Sermanet et al. [23] proposed a method based on convolutional sparse coding to pre-train Convolutional Neural Network (CNN) for pedestrian detection. In [24], Tian et al. trained extensive part detectors with weakly annotated humans to handle occlusion problems in pedestrian detection. Li et al. [25] developed a Scale-Aware Fast R-CNN framework which incorporates a large-size sub-network and a small-size sub-network into a unified architecture to capture characteristic features for detecting pedestrians of different image sizes. Zhang et al. [7] proposed a very simple but effective baseline for pedestrian detection, using an RPN followed by boosted forests on shared, high-resolution convolutional feature maps. Cai et al. [26] proposed a multi-scale deep neural network for fast and accurate pedestrian detection. Li et al. [27] made use of the dilated convolution to enhance the feature learning and construct a pedestrian detection framework along with the region proposal network and boosted decision trees. Du et al. [28] proposed a deep neural network fusion architecture applying parallel processing of multiple networks for fast and robust pedestrian detection.

Multimodal data fusion: In recent years, pedestrian detectors trained using multimodal data sources have gained popularity since multimodal sensors provide complementary information about objects of interest and lead to more robust detection results. Grassi et al. [12] proposed a novel information fusion approach to detect pedestrians, by determining the regions of interest in the video data through a lidar sensor and an infrared camera. Hwang et al. [13] have captured a large-size KAIST multispectral pedestrian dataset (KAIST) which contains well aligned RGB-thermal image pairs with dense pedestrian annotations. With this dataset, the author proposed new multispectral aggregated channel features to handle color-thermal image pairs. This reduces the average miss rate of ACF by 15%. Liu et al. [29] have designed four ConvNet fusion architectures that integrate two-branch ConvNets on different DNNs stages, all of which yield better performance compared with the baseline detector on the KAIST dataset. Xu et al. [30] designed a method to learn and transfer cross-modal deep representations in a supervised manner for the purpose of robust pedestrian detection against bad illumination conditions. However, this method is based on information of visible channel only (during the testing stage) and its performance is not comparable with ones based on multispectral data (e.g., Halfway Fusion model [29]). König et al. [31] modified the visible pedestrian detector designed by Zhang *et al.* [7] to build Fusion RPN + BDT architecture for multispectral pedestrian detection. However, the Fusion RPN + BDT model employs boosted forest for object classification which is time-consuming in both training and testing phases.

Unsupervised domain adaptation: Recently, some researches are carried out to achieve unsupervised domain adaptation in an attempt to minimize the manual annotation effort. Roth et al. [32] proposed an

unsupervised framework to learn an object detector by combining the power of a discriminative classifier with the robustness of a generative model. Based on motion analysis in video, a number of positive examples are obtained by simply checking the geometry (aspect ratio) of the motion blobs and then are fed to a discriminative classifier (AdaBoost) to learn target-related features. Wang et al. [6] proposed a new approach to automatically transfer a generic pedestrian detector to a scene-specific one in static video surveillance without manually labeling samples from the target scene. This method explores a set of context cues (e.g., locations, sizes and motions) to select high-confidence samples from the target scene to guide transfer learning. Zeng et al. [33] presented a deep model to automatically learn a scene-specific classifier and the distribution of the target samples in static surveillance videos. The specifically designed objective function not only incorporates the confidence scores of target training samples but also automatically weights the importance of source training samples. Wu et al. [34] presented a selective ensemble algorithm to select a subset of the components relevant to the target scene for recombination. The resulting model is applied for collecting high-confidence samples from unlabeled target data. It is noticed that the above-mentioned unsupervised domain adaptation methods attempt to convert a generic pedestrian detector to a scene-specific one and only demonstrate good performance on static video surveillance benchmarks (e.g., MIT-Traffic [35] or CUHK-Square [36]).

Our approach differs from the above methods distinctly by developing a unified framework which combines the auto-annotation method with a TS-RPN detector to achieve unsupervised learning of multispectral features for robust pedestrian detection. To the best of our knowledge, this is the first research work revealing that multispectral data provide complementary information to simultaneously improve detection accuracy and achieve auto-annotation of training samples. Moreover, our trained multispectral detector is not a scene-specific one. On the KAIST multispectral dataset which captures various urban scenes (dynamic viewpoints) at both daytime and nighttime, our proposed unsupervised approach using auto-annotated training data alone can achieve performance comparable to state-of-the-art DNNs-based multispectral pedestrian detectors trained using manual labels.

3. Our approach

In the visible source domain, there are many public datasets available for training pedestrian detectors, such as Caltech [8], INRIA [37], ETH [38], KITTI [39], Cityscapes [40], CityPersons [41], which are all captured during daytime under good lighting conditions. Pedestrian detectors trained on these images are sensitive to changes of illumination and weather. With the advance of multimodal sensing technology, it is possible to simultaneously capture multiple-cue information (e.g., thermal, visible, and depth) of the same scene, which provides complementary information about objects of interest and leads to more stable detection results. However, large-scale manual labeling of multispectral data captured at new target scenes is extremely time-consuming and not scalable. Therefore, the motivation of this research paper is to automatically adapt a generic pedestrian detector trained using public visible benchmarks (e.g., Caltech [8]) to new multispectral target scenes without any manual annotation efforts. This is a challenging task since pedestrian instances exhibit significantly different human-related characteristics in the visible and thermal channels during daytime and nighttime, as illustrated in Fig. 2.

In a pair of aligned visible and thermal images, a pedestrian instance is approximately located at the same position but exhibits significantly different characteristic features. Based on this important principle, we propose a novel framework for unsupervised learning of a DNN-based pedestrian detector using multispectral data. A generic pedestrian detector, pre-trained on visible images, can be automatically optimized to handle multispectral data via an unsupervised learning manner, allowing the detector to work reliably under a new extreme

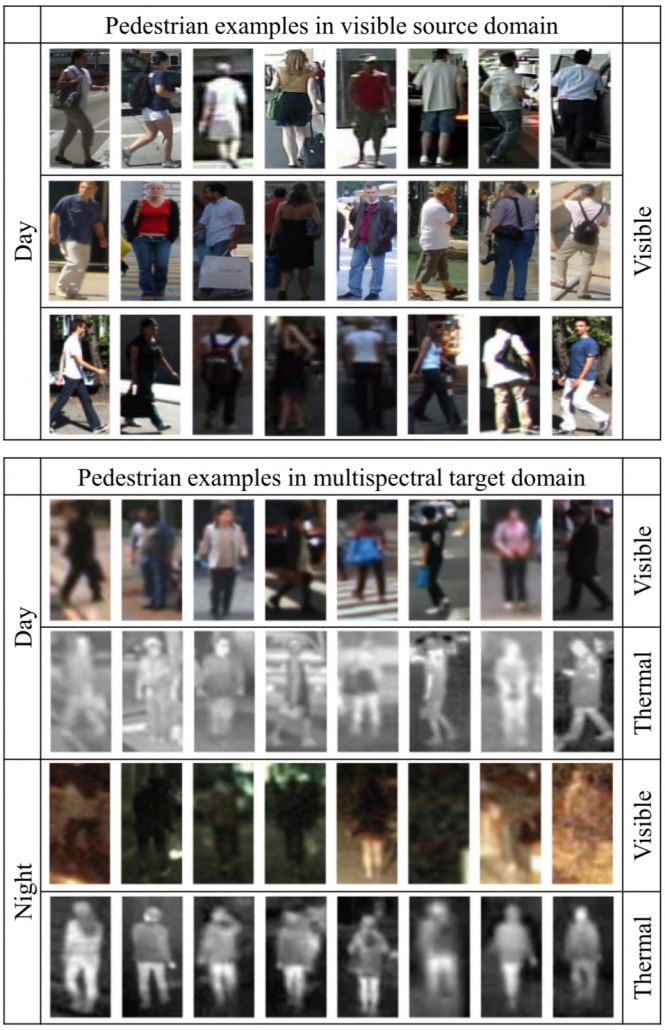


Fig. 2. Examples of pedestrians in the visible daytime source domain [8,37,39] and the multispectral (visible and thermal) target domain [13]. It is noted pedestrian instances exhibit significantly different human-related characteristics in the visible and thermal channels captured during daytime and nighttime.

condition. The diagram of our proposed method is illustrated in Fig. 3. Specifically, our approach makes use of a generic pedestrian detector pre-trained on a visible daylight dataset (the source domain) as the initial detector. Firstly, an iterative annotation approach is presented to repeatedly label pedestrian instances in visible and thermal images (the multispectral target domain). Then, a temporal tracking method is applied on visible and thermal channels to generate more confident labels and the predicted results on both channels are fused to generate the final annotations. Finally, we design a TS-RPN detector, which is trained using the automatically generated annotations. Each stream is an independent CNN with input of a visible or thermal image. The fusion of two stream features is performed at the convolutional layers, allowing our detector to effectively learn the complementary semantic features in visible and infrared channels. Some essential mathematical notations are listed in Tab. 1.

3.1. Iterative annotation

In paired color and thermal images $\{X^C; X^T\}$, pedestrians present significantly different characteristics in different image modalities, which can effectively compensate for each other. For instance, a pedestrian which cannot be clearly identified in a color image, may appear significant in its corresponding thermal image. Moreover, position consistency of objects shown in the aligned visible and infrared images

Fig. 3. The diagram of our proposed approach, which able to train a pedestrian detector using unlabeled multispectral data. Given a generic pedestrian detector trained in a visible source domain, we present a unified framework which combines the auto-annotation method with a TS-RPN detector to achieve unsupervised learning of multispectral features for robust pedestrian detection. The auto-annotation method consists of three consecutive processing steps including iterative annotation, temporal tracking and label fusion. The automatically generated annotations, together with the multispectral training images are used to train a robust multispectral TS-RPN pedestrian detector in which two-stream semantic features are fused in the convolutional layer.

Table 1
Notations used in this paper .

X	multispectral target images
$\{C,T\}$	color $\{C\}$ and thermal $\{T\}$ channels
$\{D,N\}$	daytime $\{D\}$ and nighttime $\{N\}$ scenes
Θ	pedestrian detector
θ	parameters of pedestrian detector
G	labels of pedestrian detections
b	bounding boxes
(b^x, b^y, b^w, b^h)	bounding box coordinates
v	confidence scores
E	confident labels
F	candidate labels
P	positive labels

allows us to iteratively update detection results on one channel by considering complementary information in another. Based on above considerations, a novel iterative annotation algorithm is summarized in Algorithm 1.

In Fig. 2, it is observed that human-related characteristics change significantly in the visible channel during daytime and nighttime, while pedestrians appear visually consistent in the thermal channel under different illumination conditions. Therefore, we make use of two visible detectors (daytime and nighttime) and one thermal detector to learn pedestrian features in color and thermal images, respectively. The parameters of daytime and nighttime color pedestrian detectors θ^{C_D} and θ^{C_N}, as well as the thermal pedestrian detector θ^T are all initialized using the parameters of a generic pedestrian detector θ_0, which is pre-trained using daytime visible images (the source domain) [7].

A pedestrian detector $\Theta(X;\ \theta)$ generates a number of detection results $\{G,\ b_G,\ v_G\}$, where G is the prediction label, $b_G = (b_G^x,\ b_G^y,\ b_G^w,\ b_G^h)$ denotes coordinates of its bounding box, and v_G is the confidence score. Following the same strategy used in traditional detector (e.g., support vector machine [42] or Adaboost [43]), we ignore the last softmax loss layer so that the range of confident score v_G is $[-\infty,\ +\infty]$. Here we only consider the detection results with positive confident scores. Confident labels E and candidate labels F are obtained from positive detection result G as follows:

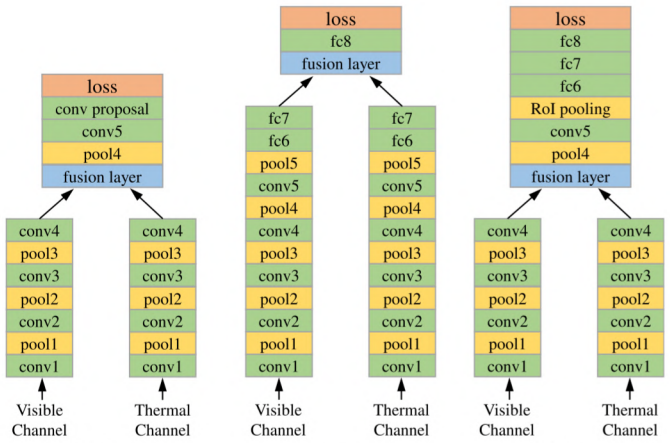
$$E = \{x \in G:\ v_x > v_{thr}\}, \tag{1}$$

$$F = \{x \in G:\ v_{thr} \geqslant v_x > 0\}, \tag{2}$$

where the high confidence threshold $v_{thr} = \mu_{v_G} + \sigma_{v_G}$, μ_{v_G} and σ_{v_G} are the mean and standard variation of all positive confident scores,

Input:
The multispectral images $X = \{X^{C_D}, X^{C_N}, X^{T_D}, X^{T_N}\}$
Parameters of a generic pedestrian detector θ_0
Initialization:
Set $\theta_0^{C_D} = \theta_0$, $\theta_0^{C_N} = \theta_0$, $\theta_0^{T} = \theta_0$ and $\widetilde{E}_0 = \{\widetilde{E}_0^{C_D}, \widetilde{E}_0^{C_N}, \widetilde{E}_0^{T_D}, \widetilde{E}_0^{T_N}\} = \emptyset$
Perform $\Theta(X^{C_D}; \theta_0^{C_D})$ and $\Theta(X^{C_N}; \theta_0^{C_N})$
Generate $E_0^{C_D}, F_0^{C_D}, E_0^{C_N}, F_0^{C_N}$ based on Eqs. 1, 2
Set $P_0^{T_D} = E_0^{C_D}$ and update θ_0^{T} to θ_1^{T} based on Eqs. 4, 5
for $k = 1, ..., K$ **do**
  // Updating visible channel detectors
    Perform $\Theta(X^{T_D}; \theta_k^{T})$ and $\Theta(X^{T_N}; \theta_k^{T})$
    Generate $E_k^{T_D}, E_k^{T_N}, F_k^{T_D}, F_k^{T_N}$ based on Eqs. 1, 2
    Update $\theta_{k-1}^{C_D}$ to $\theta_k^{C_D}$ and $\theta_{k-1}^{C_N}$ to $\theta_k^{C_N}$ based on Eqs. 3, 4, 5
    Update $\widetilde{E}_{k-1}^{T_D}$ to $\widetilde{E}_k^{T_D}$ and $\widetilde{E}_{k-1}^{T_N}$ to $\widetilde{E}_k^{T_N}$ based on Eq. 6
  // Updating thermal channel detector
    Perform $\Theta(X^{C_D}; \theta_k^{C_D})$ and $\Theta(X^{C_N}; \theta_k^{C_N})$
    Generate $E_k^{C_D}, F_k^{C_D}, E_k^{C_N}, F_k^{C_N}$ based on Eqs. 1, 2
    Update θ_k^{T} to θ_{k+1}^{T} based on Eqs. 3, 4, 5
    Update $\widetilde{E}_{k-1}^{C_D}$ to $\widetilde{E}_k^{C_D}$ and $\widetilde{E}_{k-1}^{C_N}$ to $\widetilde{E}_k^{C_N}$ based on Eq. 6
  if $(Num(\widetilde{E}_k) - Num(\widetilde{E}_{k-1}))/Num(\widetilde{E}_{k-1}) < 0.01$ **then**
    break; // Convergence condition is reached
  end
end
Output: $\{E = \widetilde{E}_K, F = F_K\}$.

Algorithm 1. Iterative annotation.



(a) TS-RPN (b) Late Fusion (c) Halfway Fusion

Fig. 4. Architectures of our TS-RPN and the state-of-the-art two-stream fusion networks (Late Fusion[45] and Halfway Fusion [29]). Green boxes represent convolutional and fully-connected layers, yellow boxes represent pooling layers, orange boxes represent loss layers and blue boxes represent fusion layers. Best viewed in color. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

respectively. Here a confident detection denotes a pedestrian instance which can be well identified using single channel information, while a candidate detection represents a possible target whose existence needs further verification by considering information from another channel. Since the pedestrian instances located at the same positions in a pair of aligned visible and thermal images, auto-annotated labels (confident labels E and candidate ones F) in one channel can be mapped to the other one to generate training samples (positive labels P). More specifically, we obtain positive labels P^{A_1} in channel A_1 by considering the latest generated labels in both channel A_1 and channel A_2 as follows:

$$P^{A_1} = \left\{ x \in E^{A_2} : \alpha < \max_{i \in \{E^{A_1} \cup F^{A_1}\}} IoU(b_x, b_i) < \beta \right\}, \tag{3}$$

where IoU is the intersection over union between two bounding boxes [44] and $0 < \alpha < \beta < 1$. Here we make use of the parameter α to promote an unconfident detection result in a channel to a positive training sample if it exhibits obvious human-related characteristic in another channel. Parameter β is set to avoid overfitting, since a well-detected instance should not be considered for detector updating. In our implementation, we empirically set $\alpha = 0.5$ and $\beta = 0.9$. Since E^{T_D} is not calculated at the initialization step ($k = 0$), we directly set $P_0^{T_D} = E_0^{C_D}$ to update the thermal channel detector θ_0^{T}. The unsigned confident labels and candidate labels are considered as *ignore regions* where we do not generate negative samples.

After generating training images with positive labels $\{X, P\}$, the parameters θ of pedestrian detector are updated by minimizing a loss function. Inspired by the multi-task learning of classification and bounding box regression [19,20], the loss of detection layer combines the cross-entropy classification loss term and the bounding box regression loss term as follows:

$$L(X, P|\theta) = -\log \psi_U(X) + \lambda \sum_{i \in \{x,y,w,h\}} smooth_{L_1}\left(b_{U^+}^{i} - \hat{b}_{U^+}^{i}\right), \tag{4}$$

where $\psi_U(X) = (\psi_{U^+}(X), \psi_{U^-}(X))$ is the probability distribution over foreground classes and background ones computed by the forward propagation of the RPN, λ is a trade-off coefficient, $\hat{b}_{U^+} = \left(\hat{b}_{U^+}^{x}, \hat{b}_{U^+}^{y}, \hat{b}_{U^+}^{w}, \hat{b}_{U^+}^{h}\right)$ denotes the regressed bounding box for foreground classes, and the robust L_1 loss function $smooth_{L_1}$ is defined in [20]. In our implementation, the coefficient λ is set to 5 as suggested by Zhang et al. [7].

At each iteration step, optimal parameters θ_k are learned by minimizing the loss function $L(X, P|\theta_{k-1})$ through the gradient descent optimization algorithm as follows:

$$\theta_{k+1} = \arg\min_{\theta_k} L(X, P|\theta_k), \tag{5}$$

After iteratively updating the parameters of detectors (θ^{C_D}, θ^{C_N} and θ^{T}) a number of times, a generic visible daytime pedestrian detector is adapted to generate accurate detection results in visible daytime, visible nighttime, and thermal all-day scenes. The newly generated confident labels E_k are used to update the existing confident label set $\widetilde{E}_{k-1}$ to $\widetilde{E}_k$ as follows:

$$\widetilde{E}_k = E_k \cup \left\{ x \in \widetilde{E}_{k-1} : \max_{i \in E_k} IoU(b_x, b_i) < \eta \right\}. \tag{6}$$

A newly generated confident label E_k will replace an existing one $\widetilde{E}_{k-1}$ if they have a large overlap ($IoU \geq \eta$), otherwise is added into the confident label set. In our implementation, we empirically set $\eta = 0.5$. With the increase of iteration times, more confident labels will be identified. However, such increase becomes insignificant when iterative annotation process continues. Once the convergence condition ($\frac{Num(\widetilde{E}_k) - Num(\widetilde{E}_{k-1})}{Num(\widetilde{E}_{k-1})} < 0.01$) is reached during iteration K, the iteration loop is stopped. The output of our iterative annotation method consists of a number of confident labels and candidate ones.

3.2. Temporal tracking

The iterative annotation approach generates pedestrian detections with confident or candidate labels solely based on information presented from single pairs of aligned visible and thermal images. Inspired by the manual labeling tool used in [8], which calculates pedestrian positions in unlabeled frames using the intermediate labeled ones through a cubic interpolation, we design a simple method to promote candidate labels to confident ones by considering temporal consistency of objects in image sequences. Given a number of consecutive image frames and the annotated labels, the temporal tracking method will promote a candidate label $F(j)$ in image frame j to a confident label $\widehat{E}(j)$ if it has large overlap with a confident label in frame $j-1$ or $j+1$ as

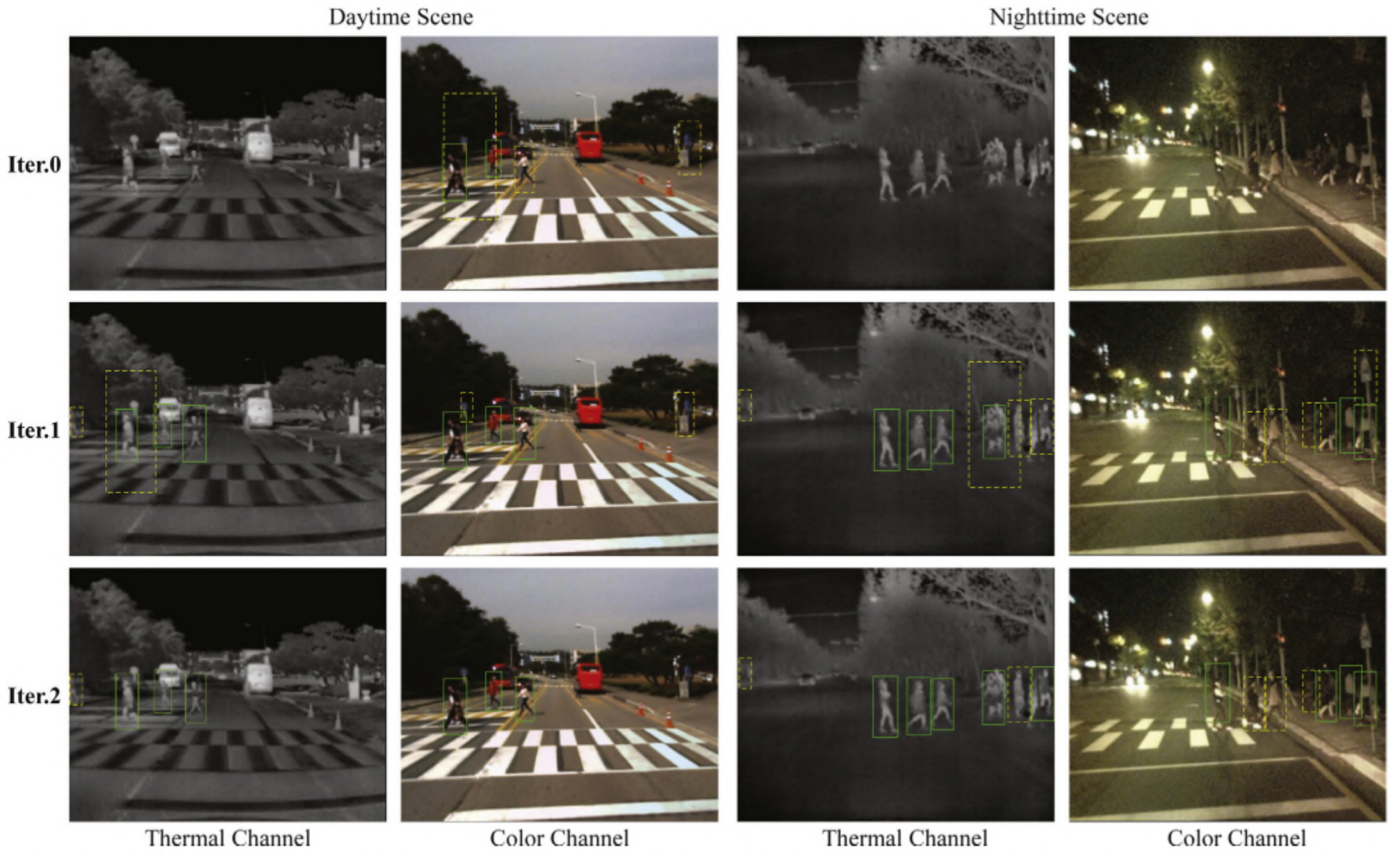


Fig. 5. Example results of iterative annotation. In the initialization step (Iteration 0), a generic RPN detector [7] well-trained in the Caltech pedestrian dataset [8] cannot correctly identify pedestrian instances when applied to multispectral images. The performance of pedestrian detection in one channel is iteratively improved by considering complementary information from another one. It is observed that more confident labels can be correctly identified and some false candidate ones are gradually removed. Note green bounding boxes show confident labels and yellow bounding boxes in dashed line show candidate ones. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

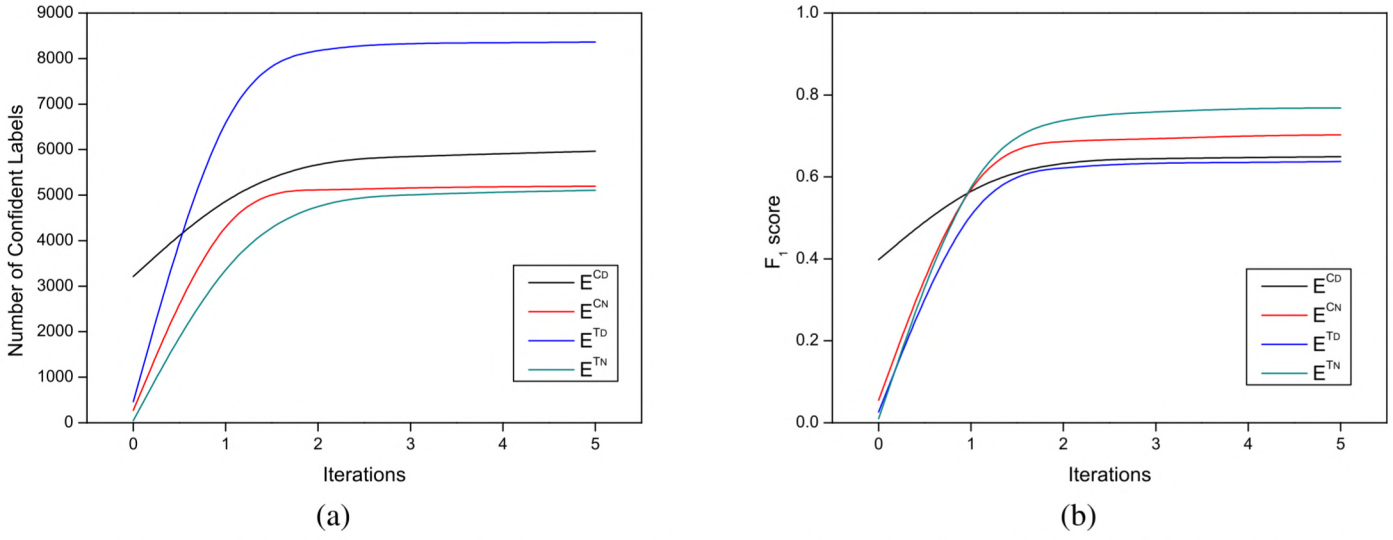


Fig. 6. The quantitative evaluation results of iterative annotation. (a) The total numbers and (b) the F_1 scores of confident labels generated in different iterations. Note Iteration 0 represents the initialization step. It is observed that the total numbers and the F_1 scores of generated confident labels both increase during the iterative annotation processes, and such improvement becomes insignificant after 5 iterations.

follows:

$$\widehat{E}(j) = \left\{ x \in F(j) : \max_{i \in (E(j-1) \cup E(j+1))} IoU(b_i, b_{F_j}) < \gamma \right\}, \tag{7}$$

where γ is a threshold parameter to impose temporal consistency of objects between adjacent frames. In our implementation we experimentally set $\gamma = 0.95$. The promoted confident labels $\widehat{E}(j)$ will be iteratively added into the existing confident label set to promote candidate labels, until there is no more $\widehat{E}_i$ generated.

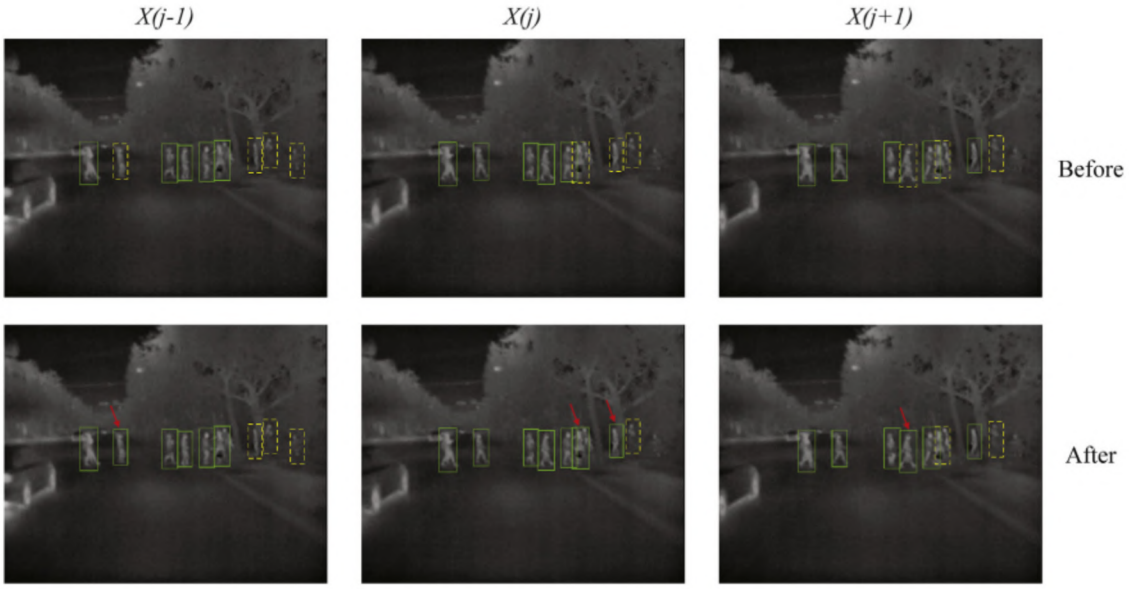


Fig. 7. Processing results of temporal tracking method on three consecutive image frames. It is noticed that a number of candidate labels have been correctly promoted to confident ones by applying the temporal tracking technique. Note green bounding boxes show confident labels and yellow bounding boxes in dashed line show candidate ones. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
The number of confident labels after temporal tracking.

	E^{C_D}	E^{T_D}	E^{T_N}	E^{C_N}
Iterative annotation	5967	8364	5104	5194
Temporal tracking	6497	8963	5697	5594

Table 3
The F_1 score of confident labels after temporal tracking.

	E^{C_D}	E^{T_D}	E^{T_N}	E^{C_N}
Iterative annotation	0.649	0.638	0.768	0.703
Temporal tracking	0.671	0.650	0.804	0.710

3.3. Label fusion

We introduce a simple scheme to merge the generated labels from two channels into a single set, allowing each pair of images have identical annotations. Given the individually generated labels $\{E^C, F^C\}$ and $\{E^T, F^T\}$ from two channels, the confident labels E^{C_D} and E^{T_N} in visible daytime and thermal nighttime (most reliable channels for pedestrian detection [13]) are directly selected as positive results. Then, the remaining confident labels will also be considered as positive samples if they also present enough human-related characteristics in the complementary channel. More specifically, a remaining confident detection and a candidate one are merged into a single one to form a positive sample, if they have a large overlap area (here we set the overlap threshold $\zeta = 0.7$). Finally, a non-maximum suppression (NMS) [8] with a threshold $\eta = 0.5$ is applied to merge the highly overlapped positive regions. The label fusion method works as follows:

$$P = P^D \cup P^N, \tag{8}$$

$$P^D = NMS\left(E^{C_D} \cup \left\{x \in E^{T_D}: \max_{i \in F^{C_D}} IoU(b_x, b_i) > \zeta\right\}, \eta\right), \tag{9}$$

$$P^N = NMS\left(E^{T_N} \cup \left\{x \in E^{C_N}: \max_{i \in F^{T_N}} IoU(b_x, b_i) > \zeta\right\}, \eta\right), \tag{10}$$

where P^D and P^N are the selected positive labels during daytime and nighttime, respectively. The unselected confident and candidate labels are considered as *ignore regions* where we do not generate negative samples. As shown in our experiments (Section 4.3.3), the label fusion approach results in more reliable detections, by recovering missing targets based on information in the complementary channel. The generated final annotations are used to train a two-stream RPN pedestrian detector, which is described next.

3.4. Two-stream RPN pedestrian detector

With a proper fusion strategy, complementary information in visible and infrared channels should lead to better pedestrian detection performance. However, most of the previous studies are reported on detecting pedestrians using either color or thermal images individually [7,23,25,46,47], rather than leveraging color and thermal images simultaneously. In this section, we present a novel TS-RPN model for pedestrian detection, extended from the state-of-the-art general object detector [19]. Motivated by the recent two-stream architecture for video recognition, where RGB image and optical flow information are encoded separately using two independent CNN streams [48]. It is straightforward to incorporate the two-stream architecture into our RPN detector, making it capable of encoding features in two image modalities.

The architecture of our TS-RPN detector is shown in Fig. 4(a). Specifically, the TS-RPN model uses two deep convolutional neural networks (DCNNs), each of which is a 16-layer VGGnet [49], to compute semantic features from both visible and infrared images. Two-stream features are fused at the forth convolutional layer (*conv*4) by simply concatenating two feature maps together, where 1×1 convolutional layer (*fusion layer*) is applied for performing linear combination of two feature maps. Our two-steam detector is closely related to the Late Fusion model [45] as shown in Fig. 4(b) and the Halfway Fusion model [29] as shown in Fig. 4(c). Different from the two fusion models, our TS-RPN avoids using the last three fully-connected networks (*fc*6, *fc*7, *fc*8), which not only reduces the size of model but also obtains significant performance improvement. More experimental results are provided in Section 4.4.

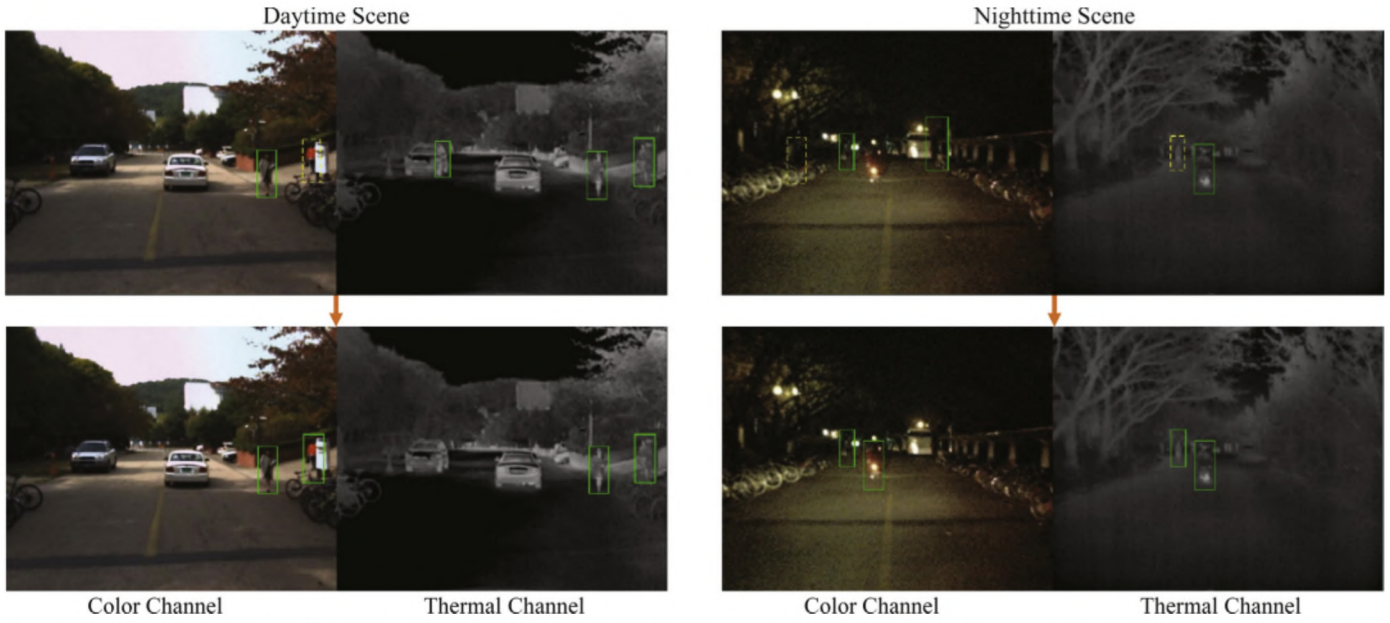


Fig. 8. Label fusion results of daytime and nighttime scenes. It is observed that more reliable detections are generated by recovering missing targets by considering information in the complementary channel. Moreover, some false detections caused by clutter background are removed since they present no/low human characteristics in either visible or thermal channel. Note green bounding boxes show confident labels and yellow bounding boxes in dashed line show candidate ones. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 4
The numbers and F_1 scores of positive labels after label fusion.

	P^D	P^N	P
Number of positive labels	7340	6027	13367
F_1 score	0.698	0.821	0.750

4. Experiments

4.1. Dataset

We evaluate the proposed unsupervised pedestrian detection framework on KAIST multispectral pedestrian benchmark [13]. In total, KAIST training dataset contains 50,172 aligned color-infrared image pairs captured at various urban locations and under different lighting conditions with dense annotations. We sample images every 2 frames and obtain 25,086 training images following [29]. The testing dataset of KAIST contains 2,252 image pairs in which 797 pairs were captured during nighttime. The original annotations under the "reasonable" setting (pedestrians are larger 55 pixels and at least 50% visible) are used for performance evaluation [13].

4.2. Implementation details

Single-scale training and testing strategy is applied without using feature pyramids [7,19,20,50]. For RPN and TS-RPN training, an anchor is considered as a positive sample if its Intersection-over-Union (IoU) ratio with one positive label is greater than 0.5, and otherwise negative. Please note that no negative samples will be obtained within the *ignore regions*. We apply the image-centric training scheme to generate mini-batches, which consist of 1 image and 120 randomly selected anchors (the ratio of positive and negative anchors is 1:5), for computing the loss [19,20]. Other hyper-parameters of RPN and TS-RPN are set as in [19]. Except in the experiment of iterative annotation, RPN and TS-RPN are initialized using the parameters of pre-trained VGG-16 model [49] on the ImageNet dataset [51], and the Conv fusion layer in TS-RPN is initialized with a Gaussian distribution. All the models are fine-tuned with Stochastic Gradient Descent [52] for the first epochs with learning rate (LR) 0.001 and one more epoch with LR 0.0001.

4.3. Evaluation of auto-annotation

We use the KAIST multispectral training dataset, which contains a large number of manual labels, to evaluate our proposed auto-annotation approach. At the end of each processing step (iterative annotation, temporal tracking, label fusion), a number of auto-annotated labels are generated. A confident/positive label is defined as a true positive (*tp*) if it can be successfully matched to a ground truth one (*IoU* of their bounding boxes exceed 50%). Unmatched confident/positive labels and unmatched ground truth are considered as false positives (*fp*) and false negatives (*fn*), respectively. Referring to Dollar et al. [8], confident/positive labels matched to ignore ground truth ones do not be counted as true positives, as well unmatched ignore ground truth labels are not considered as false negatives. The F_1 score [53], which is the harmonic mean of precision and recall, is calculated to quantitatively evaluate the performance of individual steps of the proposed auto-annotation method as follows:

$$F_1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}, \tag{11}$$

where precision and recall are computed as:

$$Precision = \frac{tp}{tp + fp}, \quad Recall = \frac{tp}{tp + fn} \tag{12}$$

4.3.1. Iterative annotation results

The proposed iterative annotation method generates a number of confident and candidate labels in thermal and color channels. In Fig. 5 we show some examples of the iterative annotation results. At the initialization step (Iteration 0), a generic RPN detector [7] well-trained in the Caltech pedestrian dataset [8] cannot produce satisfactory detection results when applied to multispectral images. It is observed that more pedestrian instances can be correctly identified and some false candidate ones are gradually removed during the iterative annotation process.

Also, we count the total number of confident labels generated during each iteration and calculate their F_1 scores. The quantitative

Fig. 9. Examples of final auto-annotated positive labels and ground truth. (a) Our auto-annotation results are comparable with ground truth; (b) our results are better than the ground truth annotations (some true positives which are missed by human observers are correctly identified by our auto-annotation approach); (c) our results are not satisfactory (some insignificant pedestrian samples are not correctly annotated). Note green bounding boxes show positive labels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

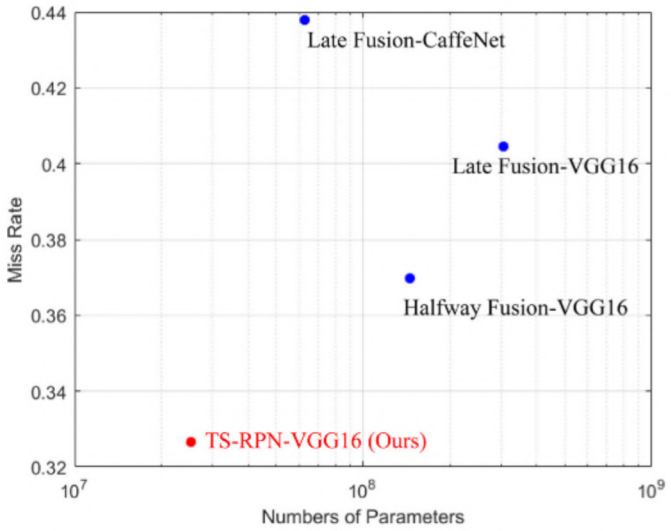


Fig. 10. Comparison of miss rate and Numbers of parameters in our TS-RPN and the state-of-the-art two-stream fusion networks (Late Fusion and Halfway Fusion).

evaluation results are shown in Fig. 6. It is observed that the total numbers and the F_1 scores of generated confident labels both increase during the iterative annotation processes. However, such increase becomes insignificant after 5 iterations. Then the convergence condition $(\frac{Num(\widetilde{E}_k) - Num(\widetilde{E}_{k-1})}{Num(\widetilde{E}_{k-1})} < 0.01)$ is reached and the iteration loop is stopped.

4.3.2. Temporal tracking results

After the step of iterative annotation, the proposed temporal tracking method is applied to generate more confident labels on visible daytime and thermal nighttime images. As shown in Fig. 7, more confident labels are promoted from candidate ones by considering temporal consistency of objects in image sequences. We count the number of confident labels before and after applying this label promotion strategy. As shown in Table 2, a significant number of candidate labels have been promoted to confident ones. Moreover, we calculate the F_1 scores and the comparative results are shown in Table 3. It is observed that higher F_1 scores, which indicate better overall precision and recall, are achieved by applying the temporal tracking technique. The results demonstrate that our proposed temporal tracking approach is a simple yet effective method to improve the quality of auto-annotated labels.

4.3.3. Label fusion results

Finally, the label fusion scheme is applied to merge the generated labels from visible and thermal channels into a single set. A pedestrian is located at the same position in a pair of aligned visible and thermal images in our final annotation result. Fig. 8 shows the label fusion results of a daytime scene and a nighttime scene, respectively. It is observed that the proposed label fusion approach results in more reliable detections by recovering missing targets based on information in the complementary channel. Moreover, some false detections caused by clutter background on the thermal channel during daytime are removed since they present no/low human characteristics in the visible channel. In Table 4, we count the total numbers and the F_1 scores of the obtained positive labels after applying the label fusion technique. Comparative results also confirm the effectiveness of our proposed label fusion scheme since higher F_1 scores are achieved.

Some final auto-annotation results are presented and compared with ground truth (manual labeling) in Fig. 9. When a pedestrian appear distinct in either visible or thermal channel as shown in Fig. 9(a), our auto-annotated results are comparable with manual labels. Moreover, our approach can accurately identify some true positives which are missed by human observers as shown in Fig. 9(b). However, some insignificant pedestrian samples are not successfully detected by our auto-annotation method, but can be still identified through human effort as shown in Fig. 9(c). The focus of our future work will be on enhancing the detection techniques to better separate object-of-interest with cluttered background. It is worth mentioning that the large-scale manual labeling of multispectral data captured at new target scenes is extremely time-consuming and not scalable. It takes around 6 seconds to manually annotate a single-channel (visible) image [8] and more than 80 hours to complete the labeling of the KAIST training dataset (containing more than 50k multispectral image pairs). In comparison, the processing time of the proposed auto-annotation pipeline (including iterative annotation for 5 iterations, temporal tracking, and label fusion) is ~ 30 hours. These auto-annotation results can be used to train a detector, making our approach readily applicable to new target scenes without extra manual labeling efforts.

4.4. Evaluation of two-stream RPN pedestrian detector

We apply the final auto-annotated labels to train our unsupervised two-stream network (U-TS-RPN), and the manual annotations in KAIST training dataset to train the supervised two-stream network(TS-RPN) for comparison. Our proposed U-TS-RPN (unsupervised) and TS-RPN (supervised) are comparing with two other multispectral pedestrian detectors Halfway Fusion [29] and ACF [13]. The detection results by the method [29] are kindly provided by its authors. The ACF detector is trained using 10-channel aggregated features to fuse color and thermal images.

In Fig. 11, we plot miss rate against false positives per image (FPPI) (using log-log plots) by varying the threshold on detection confidence. In our implementation, MR is computed by averaging miss rate at nine FPPI rates evenly spaced in log-space from the range 10^{-2} to 10^{0} [7,29].

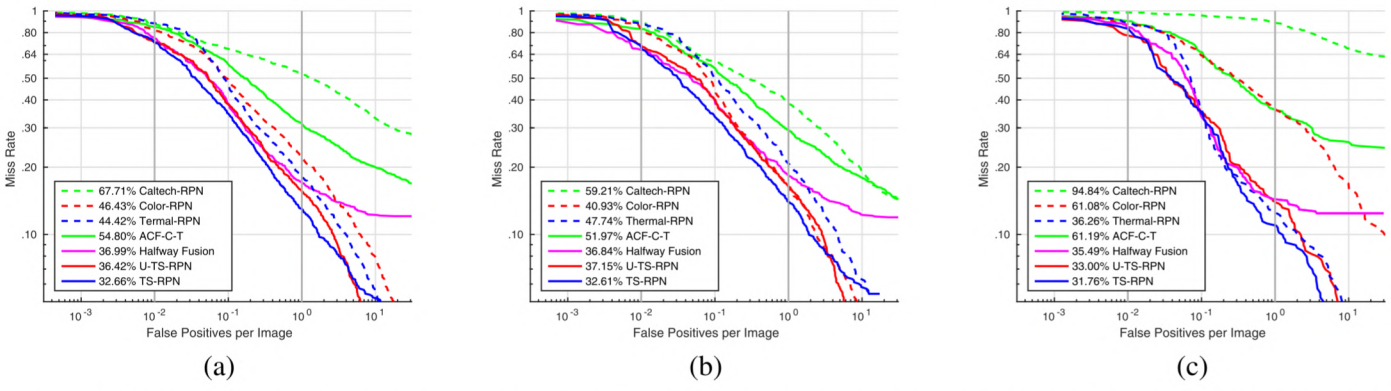


Fig. 11. Comparisons on the KAIST dataset using an IoU threshold of 0.5 to determine True Positives during all-day (a), daytime (b), and nighttime (c) (legends indicate MR).

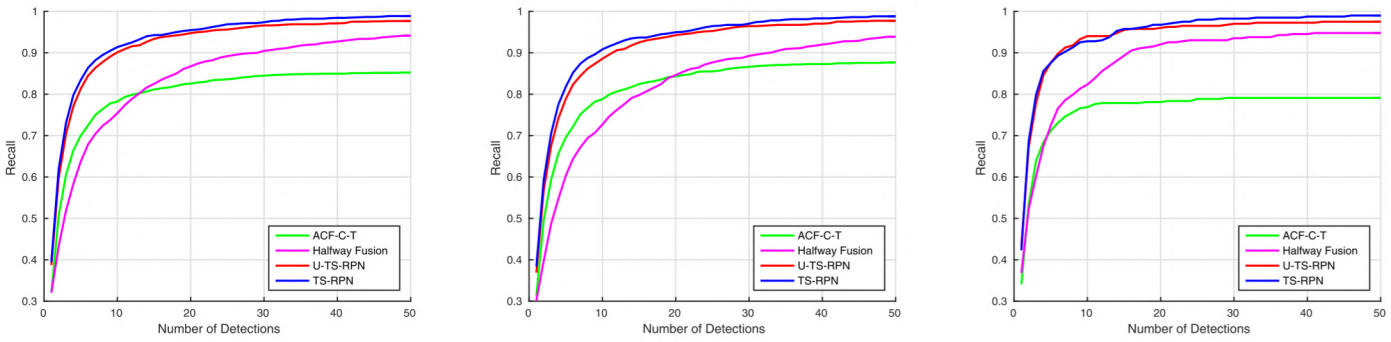


Fig. 12. Comparison of pedestrian recall vs. number of detections during all-day (a), daytime (b), and nighttime (c).

Also, we calculate the log-average miss rate (MR) to summarize performance of different detectors including Caltech-RPN (detector pre-trained on Caltech dataset [8]), Color-RPN (detector trained using visible images of KAIST dataset [13]), Thermal-RPN (detector trained using thermal images of KAIST dataset), ACF-C-T [13], Halfway-Fusion [29], TS-RPN and U-TS-RPN.

First of all, our proposed TS-RPN can better incorporate semantic features extracted on both color and thermal channels, thus leads to improved pedestrian detection results. Its overall MR is significantly lower than results of detectors trained using single color and thermal images (Color-RPN and Thermal-RPN). Compared with other feature fusion schemes, MR of TS-RPN is 4.33% lower than the results of Halfway Fusion [29] and 22.14% lower than ones of ACF-C-T [13]. Moreover, TS-RPN is a more efficient feature fusion architecture. We compare the model size of our proposed TS-RPN model with the Late Fusion model [45] and the Halfway Fusion model [29]. As shown in Fig. 10, using the same backbone DCNNs (VGG16), TS-RPN uses significantly fewer numbers of parameters compared with the state-of-art Halfway Fusion model (25.2M vs. 144.8M). We also utilize a single Titan X GPU to evaluate the computation efficiency of Halfway Fusion model and our proposed TS-RPN. The running time of TS-RPN surpasses the state-of-art Halfway Fusion method by a large margin (0.22s/image vs. 0.40s/image). Finally, but not least, our proposed U-TS-RPN, trained using auto-annotated training samples only, can achieve performance comparable to the state-of-art supervised methods using human annotation (nighttime: U-TS-RPN (33.00%) vs. Halfway Fusion (35.49%) and daytime: U-TS-RPN (37.15%) vs. Halfway Fusion (36.84%)). Furthermore, the comparative results of recall rate at IoU 0.5 vs. the number of detections are presented in Fig. 12. Given only 20 detections per image, our proposed U-TS-RPN and TS-RPN obtains 94.8% and 95.5% recall rates respectively, which are significantly higher than results of Halfway Fusion (86.7%) and ACF (82.6%).

5. Conclusion

In this paper, we present a unified framework which combines the auto-annotation method with a TS-RPN detector to achieve unsupervised learning of multispectral features for robust pedestrian detection. An effective auto-annotation approach is proposed through iteratively labeling of pedestrian instances from visible and thermal channels. It takes ~ 30 hours to complete the auto-annotation of 50k multispectral image pairs while the manual labeling process of such data requires more than 80 hours. These automatically generated labels can be used to train a detector, making our approach readily applicable to new target scenes without extra manual labeling efforts. Moreover, a two-stream region proposal network (TS-RPN) is applied to learn the semantic features from both color and thermal channels. Complementary information from multispectral data leads to more robust pedestrian detection results. It is worth mentioning that our proposed unsupervised approach using auto-annotated training data alone can achieve performance comparable to state-of-the-art DNNs-based multispectral pedestrian detectors trained using manual labels (36.42% vs. 36.99% on miss rates) on multispectral images captured at various urban scenes during both daytime and nighttime. We will make the source-code of TS-RPN and the auto-annotation results publicly available so other researches can also make use of and evaluate. However, a noticeable drawback is that some insignificant pedestrian samples which cannot be correctly identified by our auto-annotation can be still labeled through human effort. Our future work will be focused on developing more discriminative detectors to enable better separation of object-of-interest and cluttered background. Moreover, we hope to evaluate the performance of this approach under other challenging circumstances (e.g., during rainy, foggy, hazy days).

Acknowledgment

This research was supported by the National Natural Science Foundation of China (No. 51575486, No. 51605428 and U1664264) and Fundamental Research Funds for the Central Universities.

References

[1] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, IEEE Conference on Computer Vision and Pattern Recognition, (2005), pp. 886–893 vol. 1.
[2] L.V. Gool, M. Mathias, R. Timofte, R. Benenson, Pedestrian detection at 100 frames per second, IEEE Conference on Computer Vision and Pattern Recognition, (2012), pp. 2903–2910.
[3] J. Marin, D. Vázquez, A.M. López, J. Amores, B. Leibe, Random forests of local experts for pedestrian detection, IEEE International Conference on Computer Vision, (2013), pp. 2592–2599.
[4] A. Angelova, A. Krizhevsky, V. Vanhoucke, Pedestrian detection with a large-field-of-view deep network, IEEE International Conference on Robotics and Automation, (2015), pp. 704–711.
[5] J. Schlosser, C.K. Chow, Z. Kira, Fusing lidar and images for pedestrian detection using convolutional neural networks, IEEE International Conference on Robotics and Automation, (2016), pp. 2198–2205.
[6] X. Wang, M. Wang, W. Li, Scene-specific pedestrian detection for static video surveillance, IEEE Trans. Pattern Anal. Mach. Intell. 36 (2) (2014) 361–374.
[7] L. Zhang, L. Lin, X. Liang, K. He, Is faster r-cnn doing well for pedestrian detection? European Conference on Computer Vision, Springer, 2016, pp. 443–457.
[8] P. Dollar, C. Wojek, B. Schiele, P. Perona, Pedestrian detection: an evaluation of the state of the art, IEEE Trans. Pattern Anal. Mach. Intell. 34 (4) (2012) 743–761.
[9] C.-F. Lin, C.-S. Chen, W.-J. Hwang, C.-Y. Chen, C.-H. Hwang, C.-L. Chang, Novel outline features for pedestrian detection system with thermal images, Pattern Recognit. 48 (11) (2015) 3440–3450.
[10] M. Oliveira, V. Santos, A.D. Sappa, Multimodal inverse perspective mapping, Inf. Fusion 24 (2015) 108–121.
[11] L. Snidaro, J. García, J. Llinas, Context-based information fusion: a survey and discussion, Inf. Fusion 25 (2015) 16–31.
[12] A.P. Grassi, V. Frolov, F.P. León, Information fusion to detect and classify pedestrians using invariant features, Inf. fusion 12 (4) (2011) 284–292.
[13] S. Hwang, J. Park, N. Kim, Y. Choi, I. So Kweon, Multispectral pedestrian detection: benchmark dataset and baseline, IEEE Conference on Computer Vision and Pattern Recognition, (2015), pp. 1037–1045.
[14] B. Wu, R. Nevatia, Detection of multiple, partially occluded humans in a single

image by bayesian combination of edgelet part det, in IEEE Intl. Conf. Computer Vision, (2005).
[15] P.A. Viola, M.J. Jones, Robust real-time face det, Int. J. Comput. Vision 57 (2004) 137â;154.
[16] P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral Channel Features, BMVC Press, 2009.
[17] W. Nam, P. Dollár, J.H. Han, Local decorrelation for improved pedestrian detection, Advances in Neural Information Processing Systems, (2014), pp. 424–432.
[18] S. Zhang, R. Benenson, B. Schiele, Filtered channel features for pedestrian detection, IEEE Conf. Computer Vision and Pattern Recognition, (2015).
[19] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. (2016). 1–1
[20] R. Girshick, Fast r-cnn, IEEE International Conference on Computer Vision, (2015), pp. 1440–1448.
[21] X. Zhang, J. Zou, K. He, J. Sun, Accelerating very deep convolutional networks for classification and detection. IEEE Trans. Pattern Anal. Mach. Intell. 38 (2015). 1–1
[22] S. Zhang, R. Benenson, M. Omran, J. Hosang, B. Schiele, How far are we from solving pedestrian detection? IEEE Conference on Computer Vision and Pattern Recognition, (2016), pp. 1259–1267.
[23] P. Sermanet, K. Kavukcuoglu, S. Chintala, Y. Lecun, Pedestrian detection with unsupervised multi-stage feature learning, IEEE Conference on Computer Vision and Pattern Recognition, (2013), pp. 3626–3633.
[24] Y. Tian, P. Luo, X. Wang, X. Tang, Deep learning strong parts for pedestrian detection, IEEE International Conference on Computer Vision, (2015), pp. 1904–1912.
[25] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast r-cnn for pedestrian detection, IEEE Trans. Multimedia 20 (4) (2018) 985–996.
[26] Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, European Conference on Computer Vision, Springer, 2016, pp. 354–370.
[27] J. Li, Y. Wu, J. Zhao, L. Guan, C. Ye, T. Yang, Pedestrian detection with dilated convolution, region proposal network and boosted decision trees, International Joint Conference on Neural Networks, IEEE, 2017, pp. 4052–4057.
[28] X. Du, M. El-Khamy, J. Lee, L. Davis, Fused dnn: A deep neural network fusion approach to fast and robust pedestrian detection, IEEE Winter Conference on Applications of Computer Vision, IEEE, 2017, pp. 953–961.
[29] L. Jingjing, Z. Shaoting, W. Shu, M. Dimitris, Multispectral deep neural networks for pedestrian detection, British Machine Vision Conference, (2016), pp. 73.1–73.13.
[30] D. Xu, W. Ouyang, E. Ricci, X. Wang, N. Sebe, Learning cross-modal deep representations for robust pedestrian detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2017), pp. 5363–5371.
[31] D. König, M. Adam, C. Jarvers, G. Layher, H. Neumann, M. Teutsch, Fully convolutional region proposal networks for multispectral person detection, IEEE Conference on Computer Vision and Pattern Recognition Workshops, (2017), pp. 243–250.
[32] P.M. Roth, H. Bischof, Conservative learning for object detectors, Machine Learning Techniques for Multimedia, Springer, 2008, pp. 139–158.
[33] X. Zeng, W. Ouyang, M. Wang, X. Wang, Deep learning of scene-specific classifier for pedestrian detection, European Conference on Computer Vision, Springer, 2014, pp. 472–487.
[34] S. Wu, S. Wang, R. Laganière, C. Liu, H.-S. Wong, Y. Xu, Exploiting target data to learn deep convolutional networks for scene-adapted human detection, IEEE Trans. Image Process. 27 (3) (2018) 1418–1432.
[35] X. Wang, X. Ma, W.E.L. Grimson, Unsupervised activity perception in crowded and complicated scenes using hierarchical bayesian models, IEEE Trans. Pattern Anal. Mach. Intell. 31 (3) (2009) 539–555.
[36] M. Wang, W. Li, X. Wang, Transferring a generic pedestrian detector towards specific scenes, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 3274–3281.
[37] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in IEEE Conf. Computer Vision and Pattern Recognition, (2005).
[38] A. Ess, B. Leibe, L. Van Gool, Depth and appearance for mobile scene analysis, in IEEE Conf. Computer Vision and Pattern Recognition, (2007).
[39] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? The kitti vision benchmark suite, in IEEE Conf. Computer Vision and Pattern Recognition, (2012).
[40] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, IEEE Conf. Computer Vision and Pattern Recognition, (2016).
[41] B.S. Shanshan Zhang, R. Benenson Citypersons: A diverse dataset for pedestrian detection, in: arXiv:1702.05693, 2017.
[42] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (3) (1995) 273–297.
[43] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, IEEE Conference on Computer Vision and Pattern Recognition, 1 IEEE, 2001. I–I
[44] Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in IEEE Conf. Computer Vision and Pattern Recognition, (2016).
[45] J. Wagner, V. Fischer, M. Herman, S. Behnke, Multispectral pedestrian detection using deep fusion convolutional neural networks, European Symposium on Artificial Neural Networks, (2016).
[46] S.K. Biswas, P. Milanfar, Linear support tensor machine with lsk channels: pedestrian detection in thermal infrared images, IEEE Trans. Image Process. (2017).
[47] X. Zhao, Z. He, S. Zhang, D. Liang, Robust pedestrian detection in thermal infrared imagery using a shape distribution histogram feature and modified sparse representation classification, Pattern Recognit. 48 (6) (2015) 1947–1960.
[48] S. Karen, A. Zisserman, Two-stream convolutional networks for action recognition in videos, Advances in Neural Information Processing Systems, (2014), pp. 568–576.
[49] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, International Conference on Learning Representations, (2015).
[50] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, European Conference on Computer Vision, Springer, 2014, pp. 346–361.
[51] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vision (2015).
[52] N.-Y. Liang, G.-B. Huang, P. Saratchandran, N. Sundararajan, A fast and accurate online sequential learning algorithm for feedforward networks, IEEE Trans. Neural Netw. 17 (6) (2006) 1411–1423.
[53] E. Ristani, F. Solera, R. Zou, R. Cucchiara, C. Tomasi, Performance measures and a data set for multi-target, multi-camera tracking, European Conference on Computer Vision, Springer, 2016, pp. 17–35.