Pattern Recognition 80 (2018) 143–155

Contents lists available at ScienceDirect

Pattern Recognition

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

Unified multi-spectral pedestrian detection based on probabilistic fusion networks

Kihong Park, Seungryong Kim, Kwanghoon Sohn*

The School of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea

ARTICLE INFO

Article history:
Received 10 May 2017
Revised 19 January 2018
Accepted 4 March 2018
Available online 13 March 2018

Keywords:
Multi-spectral sensor fusion
Pedestrian detection
Channel weighting fusion
Probabilistic fusion

ABSTRACT

Despite significant progress in machine learning, pedestrian detection in the real-world is still regarded as one of the challenging problems, limited by occluded appearances, cluttered backgrounds, and bad visibility at night. This has caused detection approaches using multi-spectral sensors such as color and thermal which could be complementary to each other. In this paper, we propose a novel sensor fusion framework for detecting pedestrians even in challenging real-world environments. We design a convolutional neural network (CNN) architecture that consists of three-branch detection models taking different modalities as inputs. Unlike existing methods, we consider all detection probabilities from each modality in a unified CNN framework and selectively use them through a channel weighting fusion (CWF) layer to maximize the detection performance. An accumulated probability fusion (APF) layer is also introduced to combine probabilities from different modalities at the proposal-level. We formulate these sub-networks into a unified network, so that it is possible to train the whole network in an end-to-end manner. Our extensive evaluation demonstrates that the proposed method outperforms the state-of-the-art methods on the challenging KAIST, CVC-14, and DIML multi-spectral pedestrian datasets.

1. Introduction

Pedestrian detection has been a vital problem of machine learning due to its paramount relevance in commercial systems, spanning from self-driving cars to autonomous surveillance [1]. Recent advances in color sensor technologies and learning-based detection algorithms have encouraged the performance in several scenarios. However, a pedestrian may appear under varying conditions of illumination, weather, resolution, and occlusions. These restrictions make pedestrian detection in a color image more challenging.

Most efforts on traditional pedestrian detection using color sensors have focused on two key components: feature representation and classifier. The first one is the design of a feature to encode pedestrian characteristics reliably. Conventionally, thanks to the great success of histograms of oriented gradients (HOG) [2] and aggregated channel features (ACF) [3], many other variants and combinations have been proposed in the last decade, such as integral channels [4]. The second main component is the design of the classifier. For this, support vector machines (SVM) [5] is the most popular choices because of their theoretical guarantee, extensibility, and good performance. Random forest ensembles [6] have also been used as an alternative type of classifier for pedestrian detection. However, an insurmountable gap exists between these handcrafted methods and human perception ability in pedestrian detection. More recently, there has been an explosion of pedestrian detection models based on convolutional neural networks (CNNs) [7]. Owing to their high capacity to represent discriminative features and distinguish the pedestrian from background clutter [8], CNN-based approaches have achieved substantially promising results compared to the state-of-the-art approaches, such as region-based CNN (R-CNN) [9], spatial pyramid pooling network (SPP-nets) [10], Fast R-CNN [11], Faster R-CNN [12], and multi-scale CNN (MS-CNN) [13].

In parallel to all these works, there is a relatively unexplored area in the field of pedestrian detection, i.e., multi-spectral sensor fusion. Multi-spectral fusion approaches supplement the data of color images with complementary information obtained from other spectral sensors. The thermal (i.e., long-wavelength infrared) camera has been one of the promising choices as it encodes the temperature information in complex scenarios such as background clutter or lack of illumination. Since ambient lighting has little effect in thermal imaging, the thermal camera has been widely used in face recognition [14], and action recognition [15]. With regard to pedestrian detection, the thermal image usually presents clear silhouettes of human objects [16,17], and can thus help boost pedestrian detection performance. However, except for very recent ef-

* Corresponding author.
E-mail addresses: khpark7727@yonsei.ac.kr (K. Park), srkim89@yonsei.ac.kr (S. Kim), khsohn@yonsei.ac.kr (K. Sohn).

https://doi.org/10.1016/j.patcog.2018.03.007
0031-3203/

forts [18,19], multi-spectral pedestrian detection has not been studied thoroughly. It is still an open question that how color and thermal images could be fused appropriately to obtain an optimal synergy.

In this paper, we design a unified CNN architecture to fuse color and thermal information in an end-to-end manner where complementary information can boost pedestrian detection performance even under challenging real-world environments. A key observation that emerges from our analysis of multi-spectral pedestrian detection is that if either color and thermal image is significantly degraded, the half-way fusion method in [18] compromises the detection capability between color and thermal images, and thus this method cannot guarantee an optimal performance. To address this problem, we consider all detection probabilities from color, thermal, and color-thermal fusion channels in a unified deep CNN framework. Channel weighting fusion (CWF) and accumulated probability fusion (APF) layers are introduced to selectively fuse information from different modalities at a proposal-level. Importantly, our system fully integrates APF and CWF within a single network, making it possible to train the whole network end-to-end with a standard back-propagation algorithm. We demonstrate the effectiveness of the proposed approach in several challenging multi-spectral pedestrian benchmarks including KAIST [20], CVC-14 [21], and our DIML. The main contributions of this work are summarized as follows:

- We propose a unified CNN architecture for the task of multi-spectral pedestrian detection and formulate the whole network to be learned in an end-to-end manner.
- Unlike existing multi-spectral fusion techniques [18], we comprehensively utilize color, thermal, color-thermal fusion features to maximize detection performance by synergistically using their detection probabilities with channel weighting fusion (CWF) and accumulated probability fusion (APF).
- The proposed system significantly reduces the missing rate of baseline method [18] by 5.60%, yielding a 31.36% overall missing rate on the KAIST multi-spectral pedestrian benchmark [20].

This manuscript extends its preliminary conference version [19] with the following major differences: (1) We reformulate previous shallow modules, such as region proposal detection and classification modules, as deep architectures to learn an optimal feature representation in an end-to-end manner; (2) To suppress false positives, the probability accumulation scheme is extended by leveraging a spatial similarity among neighboring proposals; (3) an extensive comparative study of the proposed method is performed with various datasets qualitatively and quantitatively.

The remainder of this paper is organized as follows. Section 2 describes related works on pedestrian detection. Section 3 provides the motivation for our work. We present the proposed method in Section 4. Experimental results and discussions are provided in Section 5, and we conclude the paper in Section 6.

2. Related work

This section describes related works on pedestrian detection methods, including deep networks for object or pedestrian detection and multi-spectral fusion approaches for pedestrian detection or other computer vision tasks.

Deep networks for object detection Traditional methods for object detection based on handcrafted features, such as HOG [2], ACF [3], and integral histogram [4], combined with shallow machine learning schemes, such as SVM [5] and random forest [6], have shown limited capacities to provide reliable detection performances. Over the past few years, deep convolutional neural networks (CNNs) based approaches have become increasingly popular owing to their reliability in object detection tasks. One of the representative studies is R-CNN [9], which utilizes convolutional activation features as in [22] to localize the object among object proposal candidates. However, it performs a CNN forward pass for each object proposal independently, and hence its computation is very slow. To overcome these limitations, SPP-nets [10] and Fast R-CNN [11] were proposed to speed up R-CNN by sharing the computation on convolutional features. Especially, Fast R-CNN [11] proposed a region-of-interest (RoI) pooling scheme that reshapes intermediate features as the desired proposal size. However, these methods showed a region proposal computation is a bottleneck. To address this problem, Faster R-CNN [12] proposed a region proposal network (RPN) that shares full-image convolutional features with the detection network, thus enabling nearly cost-free region proposal estimation. However, these methods have disadvantages in that they cannot detect a small-sized object due to the collapsing bin problem associated with RoI pooling. To solve this, Cai et al. [13] proposed a method that repeatedly performs the detection process at multiple levels of features. Gidaris and Komodakis(2015) [23] introduced a bounding box voting (BBV) scheme that refines bounding boxes by leveraging the neighboring proposals, and this scheme is only considered at bounding box re-localization process.

Deep networks for pedestrian detection In the field of pedestrian detection, deep models have been formulated to represent pedestrian-specific features. One of the pioneering works is multi-stage unsupervised feature learning [24] that automatically learns hierarchical features with unsupervised sparse auto-encoder. Li et al. [25] proposed a scale gate function to solve scale problems by capturing different characteristic features as image sizes. Tian et al. [26] solved the partial occlusion problem by considering the semantic attributes of people and scenes. While the techniques mentioned above provide reliable detection performance to some extent, none of them are designed to deal with bad lighting conditions; hence, their performance degrades in challenging scenarios, especially in nighttime situations.

Multi-spectral fusion based computer vision tasks To solve the inherent limitations of a color camera such as bad visibility and sensitivity to noise in pool lighting conditions, multi-spectral fusion approaches have been popularly used to supplement the data provided in a color image in various applications. Thanks to the success of multi-spectral image registration techniques [27,28], multi-spectral information is well combined, and various multi-spectral applications have been suggested, by using color and near-infrared (NIR) images, or color and thermal images. Feng et al. [29] proposed an image dehazing approach by modeling a dissimilarity between color and NIR images. The NIR image was also used as a guidance image in image denoising applications [30].

Multi-spectral fusion for pedestrian detection Since a pedestrian reveals distinct properties related to temperature information in a thermal image, the color and thermal image combination can enable us to overcome the inherent limitations in pedestrian detection. Recently, the KAIST benchmark [20] facilitated the study of pedestrian detection in a large-scale multi-spectral dataset, where it provided aligned color and thermal image pairs obtained using a beam splitter technique. The CVC-14 benchmark [21] also provided continuous video information of color and thermal images. Multispectral ACF (MACF) [20], as a pioneering work of multi-spectral (i.e., color and thermal) fusion based pedestrian detection, has shown that additional thermal information could help to highly boost the pedestrian detection performance, especially in night environments. However, this method showed limitations due to their handcrafted features that do not have enough discriminative power to represent a pedestrian in challenging circumstances. Recently, Liu et al. [18] proposed a convolutional network to fuse color and thermal images as a variant of Faster R-CNN [12]. However, their fusion technique that extracts fused features from both

(a) (b) (c) (d)

Fig. 1. Examples of complementary detection results using convolutional features from only color, thermal, and color-thermal fusion channels. Boxes with yellow and green color represent ground-truth and estimated bounding boxes, respectively. With a thermal image in (a), the pedestrian candidates are detected using individually learned features on (b) the color image only, (c) the thermal image only, and (d) the existing fusion feature [18] on both color and thermal images. The existing fusion method [18] cannot fully reveal the complementary potential between color and thermal images, but rather individually learned features on each modality can produce reliable performances in some circumstances. By synergistically fusing these features, our system can maximize the pedestrian detection performance under all challenging circumstances. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(a) (b)

Fig. 2. Examples of pedestrians in multi-spectral color and thermal images. (a) When the thermal image is saturated due to sunlight or ambient temperature, a color image can be a reliable cue to detect a pedestrian. (b) However, when the color image is degraded in bad lighting conditions, thermal image can contribute to detecting a pedestrian robustly. By jointly leveraging both color and thermal images, our approach overcomes these challenging limitations.

color and thermal images compromises the detection performance; hence, they cannot guarantee an optimal performance. Unlike this, our method estimates detection probabilities from not only such a fused feature but also each distinctive feature from each modality simultaneously.

3. Motivation

Traditional methods such as half-way fusion strategy [18] use separate sub-networks to extract the features for each modality and combine them through an additional fully-connected layer to extract complementary pedestrian features for both modalities. In this paper, we argue that this strategy may not fully reveal the complementary potential between color and thermal images. The half-way fusion [18] cannot cover diverse situations, as shown in Fig. 1. It fails to detect pedestrians as in Fig. 1(d). Conversely, these pedestrians are detectable by considering the single image modality only, which is shown in Fig. 1(b) and (c). These examples illustrate the fact that the half-way fusion method [18] is not sufficient to achieve the best multi-spectral pedestrian detection synergy. In some cases, the individually learned network for each modality can provide reliable performance compared to the fusion-based approach. Furthermore, we observed that if either the color or thermal image is significantly degraded, the detection performance of the half-way fusion method [18] is compromised, thus providing limited performance. We summarize such challenging conditions as follows:

- The thermal image is occasionally saturated by sunlight or ambient temperature. (see Fig. 2(a))
- Under bad lighting conditions, such as nighttime, the color image has low visibility and is vulnerable to noise. (see Fig. 2(b))

Nevertheless, we still benefit from the half-way fusion method [18] when both color and thermal images provide useful information in depicting pedestrians. Based on the discussion and observation, we propose to consider all pedestrian probabilities from not only each modality but also the half-way fusion method in a unified deep CNN framework. To realize this, our method focuses on fusing the three-branch information in a boosting manner.

4. Unified multi-spectral pedestrian detection

4.1. Problem formulation and overview

Given a color image I and a thermal image F, the objective of our pedestrian detection system is to localize the location and scale of a pedestrian robustly even under challenging conditions such as bad visibility or noise during nighttime. Formally, for each pixel $i = [i_{\mathbf{x}}, i_{\mathbf{y}}]^T$, the pedestrian proposal candidate (i.e., location and scale) and the pedestrian probability are first estimated, and the pedestrian is then detected in an image domain. Our approach belongs to sensor fusion approaches, and uses color and thermal images jointly to leverage the complementary information from each modality synergistically. Unlike conventional fusion schemes, e.g., half-way fusion strategy [18], that utilize only common pedestrian features from each modality, our method is formulated to utilize not only such a fused feature but also each distinctive feature from each modality simultaneously.

By leveraging CNNs [22], our pedestrian detection method is formulated as three sub-networks, including feature extraction network, region proposal network, and inference network, as shown in Fig. 3. Concretely, the feature extraction network is designed to extract distinctive features of pedestrians from each modality, which can be encoded on a color and thermal image independently. To boost the detection performance, we also extract a fused feature that encodes complementary information between each modality through the concatenation of intermediate convolutional activations. Based on the observation that these three kinds of features can encode each distinctive characteristic and provide different robustness depending on environments, the pedestrian probabilities are estimated from all convolutional features to fully utilize complementary information from each modality. In the region proposal network, the region proposal generation is formulated on the fused pedestrian feature. Furthermore, we propose a channel weighting fusion (CWF) layer to determine an optimal pedestrian feature among three-branch features at each proposal. In the inference network, we propose an accumulated probability fusion (APF) layer. To suppress false positives and estimate optimal probabilities

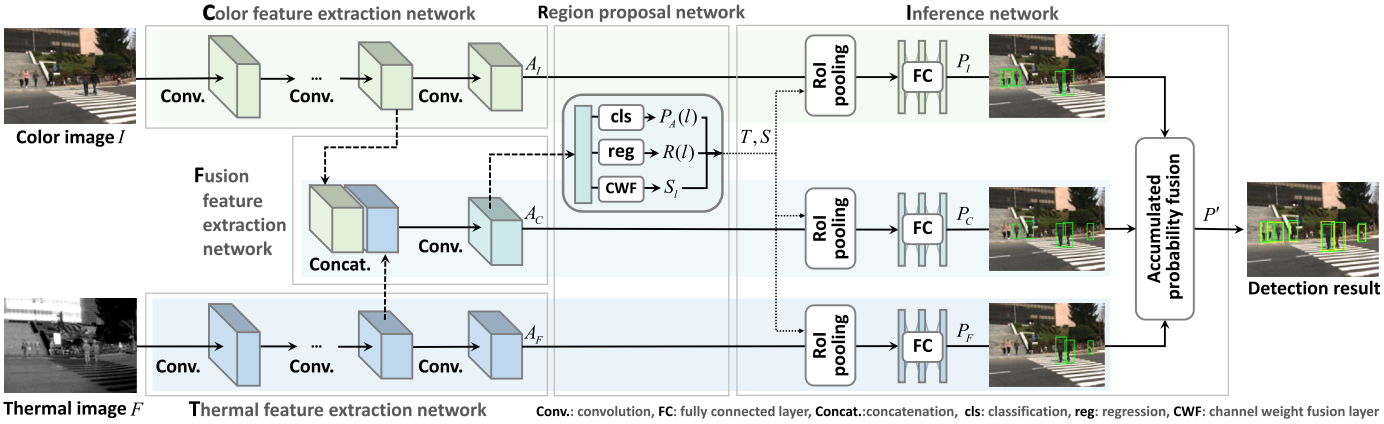


Fig. 3. The illustration of our overall pedestrian detection framework. Our proposed network takes a color-thermal image pair as inputs and produces the pedestrian location as outputs. Our proposed network consists of three sub-networks, including a feature extraction network, region proposal network, and inference network. In the resultant images, boxes with yellow and green color represent ground-truth and estimated bounding boxes, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

from each feature, the ACF layer accumulates the pedestrian probabilities of each feature with neighboring bounding boxes at the proposal-level. Since our network consists of fully convolutional modules, it can be learned in an end-to-end manner without any approximation or handcrafted processes.

4.2. Network architecture

In this section, we first describe the details of our pedestrian detection method, consisting of three sub-networks, namely feature extraction network, region proposal network, and inference network, and then show how they can be learned in an end-to-end manner.

4.2.1. Feature extraction network

Compared to previous methods detecting pedestrians in an image only [12], pedestrian detection using both color and thermal images can provide an outstanding performance even under challenging conditions such as night environments. Inspired by this, some approaches have tried to utilize both color and thermal images to detect pedestrians (e.g., half-way fusion [18]). However, as exemplified in Fig. 1, the fused pedestrian feature extracted from both color and thermal images cannot guarantee consistent robust performance in all situations.

Based on these observations, we propose three-branch feature extraction networks, consisting of color feature extraction network, thermal feature extraction network, and fused feature extraction network. By independently learning the pedestrian feature from each modality, our pedestrian detection method can extract distinctive properties of each modality while benefiting from the complementary property of each modality through the fused pedestrian feature. The convolutional features estimated from each network, denoted as A_I, A_F, and A_C, are used to estimate pedestrian probabilities simultaneously, as described in the following section.

Similar to existing detection methods [12], each feature extraction network is formulated as successive convolutions. For the fused feature extraction network, we first extract the intermediate convolutional activations from color and thermal feature extraction networks and then concatenate them. For a fusion feature, we utilize the convolutional activations of 'Conv4-3' due to their robustness and computational efficiently. Moreover, we also build additional convolutional layers with the same size as 'Conv5' after the fusion concatenation and Network-in-Network (NIN) layer [18]; thus, A_I, A_F, and A_C have the same size of feature dimension.

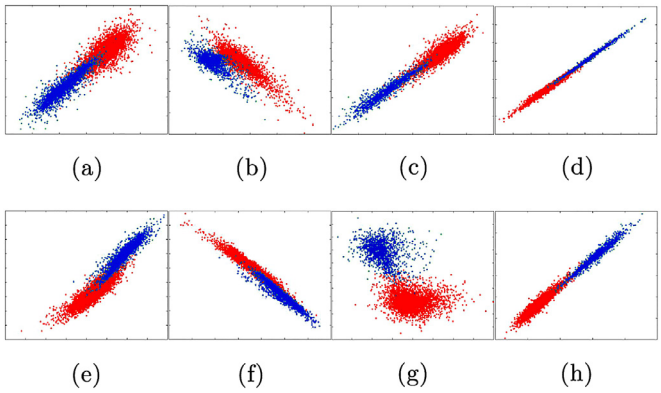


Fig. 4. Visualization of convolutional features using linear discriminant analysis (LDA), extracted on (b) color A_I, (c) thermal A_F, and (d) fusion channel A_C in comparison to (a) existing handcrafted feature (i.e., ACF [3]) under (top) day and (bottom) night environments. Red and blue dots indicate negative and positive pedestrian samples, respectively. Compared with ACF [3], the state-of-the-art handcrafted method, the proposed convolutional features are more discriminative to reliably distinguish the pedestrian and non-pedestrian patches. Furthermore, three channel features have shown different distributions depending on day and night environments, and thus an effective fusion technique for these convolutional features is essential. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Fig. 4 visualizes the distinctive and complementary properties of convolutional features in our network in comparison to the handcrafted method, i.e., ACF [3], without any techniques such as hard negative mining scheme or jittering the data, on the KAIST multi-spectral benchmark [20]. Fig. 4 (b)–(d) shows examples of linear discriminant analysis (LDA) of each convolutional feature, A_I, A_F, and A_C, to distinguish the pedestrian and non-pedestrian candidates. We extracted these convolutional features on pedestrian region proposals as outputs of the region proposal network, which will be described in details in the following section. From these case studies, we can observe two kinds of intuitions. First, compared to handcrafted features such as ACF [3] that cannot distinguish the pedestrian reliably, our convolutional features are more discriminative even under challenging day and night environments, owing to their deeper architectures. Second, according to environments such as day and night, each convolutional feature has different properties and robustness. There is no single feature to provide consistently reliable performance for all situations. Thus, an effec-

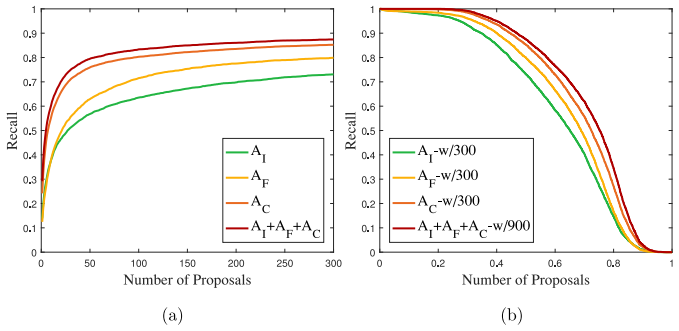


Fig. 5. Comparison of pedestrian detection performance using convolutional features by varying the number of proposals, including (a) recall vs. the number of proposals and (b) recall vs. intersection of union (IoU). Since the performance gap between results using all three channels ($A_I + A_F + A_C$) and results using only fused channel (A_C) is marginal, our method only utilizes the fused convolutional feature A_C with 300 proposals.

tive fusion technique of these convolutional features is essential to maximize the detection performance.

4.2.2. Region proposal network

Based on the estimated convolutional features, the pedestrian candidates can be hypothesized using an additional network to generate pedestrian proposal candidates. Similar to [18], we adopt the region proposal network (RPN) that takes convolutional features as inputs and outputs a set of rectangular proposals parameterized by a tuple $T(i) = [i_{\mathbf{x}}, i_{\mathbf{y}}, \nabla i_{\mathbf{x}}, \nabla i_{\mathbf{y}}]^T$, where $(i_{\mathbf{x}}, i_{\mathbf{y}})$ is the center point and $(\nabla i_{\mathbf{x}}, \nabla i_{\mathbf{y}})$ are horizontal and vertical diameters of the bounding box, each with an anchor probabilities $P_A(i)$, where anchor means the pre-defined reference bounding box. We also model this process with successive convolutions. Even though all three-branch convolutional pedestrian features, A_I, A_F, and A_C, can be used for the region proposal network, our method only utilizes the fused convolutional feature A_C.

Fig. 5 shows the pedestrian proposal detection performance on the KAIST multi-spectral benchmark [20]. In these case studies, we extracted 300 pedestrian proposals from each channel feature. As shown in the statistics, the accuracy of 900 pedestrian proposals consisting of 300 proposals from each channel was similar to the accuracy of 300 proposals from the fusion channel only. Thus, we observe that the capacity of pedestrian proposal detection saturates at 300 proposals, and there is no need to extract proposal candidates from all features A_I, A_F, and A_C; therefore, we utilize A_C as an input for the region proposal network.

To generate pedestrian region proposals, we then slide a small network over the feature A_C. This small network takes a 3×3 spatial window of the input convolutional feature map, and each sliding window is mapped to a lower-dimensional feature. It is then fed into three sibling fully connected layers for box classification loss, box regression loss, and channel weighting fusion loss, which will be explained in the following section.

To detect reliable pedestrian bounding boxes among anchors, we train a regression model for each pedestrian bounding box $T(i) = [i_{\mathbf{x}}, i_{\mathbf{y}}, \nabla i_{\mathbf{x}}, \nabla i_{\mathbf{y}}]^T$. Since it is difficult to directly regress this with the ground truth bounding box $T^*(i) = [i^*_{\mathbf{x}}, i^*_{\mathbf{y}}, \nabla i^*_{\mathbf{x}}, \nabla i^*_{\mathbf{y}}]^T$, our network is instead learned to minimize the difference of transformations between $T(i)$ and $T_A(i)$ and those between $T^*(i)$ and $T_A(i)$, where $T_A(i)$ is the pedestrian bounding box of the pre-defined anchor, which will be described in the following section.

Channel weighting fusion (CWF) layer In our system, at each pedestrian proposal, three kinds of probabilities can be estimated from A_I, A_F, and A_C. As described earlier, the three probabilities have complementary information of each feature extraction network. To boost the detection performance, we propose a channel weighting fusion (CWF) layer as a gating function that determines which channel can maximize the detection performance at each proposal candidate. This process can be implemented as a classification problem, where for each proposal candidate $T(i)$ we learn the network to choose the label $S \in \{S_I, S_F, S_C\}$. As shown in Fig. 3, the CWF layer is also located within the region proposal network, and hence, only A_C is used to estimate S. Since A_C includes complementary information of color and thermal channels, the channel selection based on A_C is enough to provide optimal performance, which will be verified in the Section 5.2 experimentally.

Intuitively, the class label S from the CWF layer is desired to satisfy the following constraints. For a color image having high-quality visibility, the class label S_I is chosen to maximize the detection performance. Furthermore, for thermal image encoding a high distinctive property of a pedestrian at nighttime, the class label S_F is chosen. When both color and thermal images have good visibility for pedestrian detection, the class label S_C is chosen.

4.2.3. Inference network

Followed by the convolutional features on each pedestrian proposal, we design the inference network to determine the pedestrian probability and location at each pedestrian proposal. To maximize the pedestrian detection performance, we adopt the region-of-interest (RoI) pooling scheme [11], followed by sequential fully-connected layers for determining the pedestrian. To fuse the probabilities as outputs from each channel, we propose an accumulated probability fusion (APF) layer that maximizes the detection performance by considering the probabilities of neighborhood proposal candidates. It should be noted that our proposed network is formulated with a fully convolutional architecture and can therefore be learned in an end-to-end manner.

Region-of-interest (RoI) pooling Since each pedestrian convolutional feature is computed in an image domain and each pedestrian proposal has a different size, each convolutional feature at pedestrian proposals should be resized to a fixed size to be applied to the fully-connected layer. Inspired by [11], we convert the features inside any valid RoI into a small feature map with a fixed spatial extent of $H \times W$, where H and W are layer hyper-parameters independent of any particular RoI. Since our network design is based on the VGG-Net model [31], these parameter are set to be compatible with the following fully connected layer. For each RoI, the pedestrian probability from each channel can be determined as $P_I(i)$, $P_F(i)$, and $P_C(i)$.

Accumulated probability fusion (APF) layer One of the major problems encountered in the pedestrian detection task is false positives, where the pedestrian-like features from non-pedestrian proposals are erroneously detected as pedestrians. These false positives reduce the overall detection accuracy and further produce false alarms in intelligent vehicle systems. In pedestrian detection, the false positives frequently appear within two kinds of bounding box cases, including partially overlapped boxes and pedestrian-like boxes. First, the pedestrian probabilities of partially overlapped boxes are higher than those of other proposals since they partially contain pedestrian parts, which will be false positives, as exemplified in Fig. 6. As in [12], a simple non-maximal suppression (NMS) process cannot guarantee satisfactory results under these circumstances. Second, for visually pedestrian-like boxes, false positives frequently appear at an isolated region, as also seen in Fig. 6.

These two kinds of problems can be overcome by considering the correlation of its neighborhood proposals. Most existing detection methods [12] consider each pedestrian proposal independently, and thus they are sensitive to false positives. To overcome this, we propose an accumulated probability fusion (APF) layer that accumulates the estimated probabilities at each pedestrian proposal with its neighboring proposals. To realize this, the neighborhood system $N(i)$ is first defined for each pedestrian proposal $T(i)$,

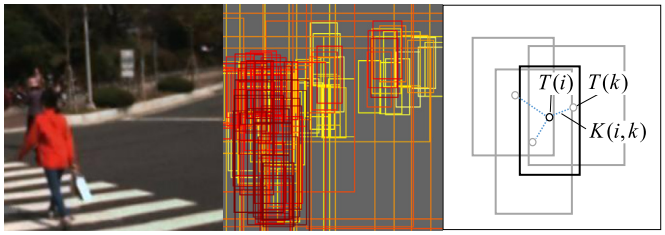


Fig. 6. Examples of an image to produce false positives in (a) and accumulated pedestrian probabilities on the image. In (b), red color represents high pedestrian probabilities and yellow color represents low pedestrian probabilities. With proposal-wise accumulation processes in an accumulated probability fusion (APF) layer using neighboring proposal candidates as in (c), the false positives can be suppressed effectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

and $T(k)$ for $k \in N(i)$ are neighboring pedestrian proposals determined by considering a spatial similarity $K(i,k) = \|T(i) - T(k)\|^2$. With these neighboring pedestrian proposals, the pedestrian probability of each candidate can be updated by accumulating the probabilities of neighboring proposals as

$$P'(i) = \frac{1}{N_P} \sum_{k \in N(i)} \sum_{l \in \{I,F,C\}} \exp(-K(i,k)) S_l(k) P_l(k), \tag{1}$$

where $P_l(k)$ is the pedestrian probability estimated at neighboring pedestrian proposals $T(k)$ and the normalization factor is $N_P = \sum_{k \in N(i)} \sum_{l \in \{I,F,C\}} \exp(-K(i,k)) S_l(k)$. It is designed to make spatially-similar proposals contribute more to estimating a final pedestrian probability with label l for each channel, reducing the possibility of false positives. Note that, similar to our scheme, Gidaris and Komodakis [23] proposed a bounding box voting (BBV) scheme that also considers the neighboring proposals. However, it only focuses on the re-positioning of the bounding boxes, leading to a marginal improvement in detection accuracy, which will also be discussed in experiments. On the other hand, our APF layer focuses on re-scoring pedestrian probabilities from the neighboring proposals, and improves detection accuracy by eliminating the false positives.

Differentiability of APF layer For end-to-end learning of the proposed system, the derivatives for the ACF layer much be computable, so that the gradients of the final loss can be back-propagated to the feature extraction network and region proposal network. Since our ACF layer is formulated as a linear summation, its derivative can be easily computed. Specifically, the derivative of the final loss $\mathcal{L}$ with respect to $P_l(k)$ can be formulated as

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial P_l(k)} &= \frac{\partial \mathcal{L}}{\partial P'(i)} \frac{\partial P'(i)}{\partial P_l(k)} \\ &= \frac{\partial \mathcal{L}}{\partial P'(i)} \frac{1}{N_P} \exp(-K(i,k)) S_l(k). \end{aligned} \tag{2}$$

This enables us to learn our network in an end-to-end manner. Table 1 represents the proposed network configuration.

4.3. Network training

To learn our overall network, consisting of a feature extraction network, region proposal network, and inference network simultaneously, we employ two loss functions for the region proposal network and inference network. Each loss function consists of bounding box classification loss and bounding box regression loss with respect to ground truth pedestrian bounding boxes. Furthermore, to learn the CWF layer, we employ an additional classification loss. Similar to other pedestrian detection methods [12], our training procedure needs the ground truth pedestrian bounding box T^*. To summarize, the total loss function is formulated as a multi-task loss as follows:

$$\mathcal{L}_{total} = \mathcal{L}_{RPN}(T, P_A, S, T^*, P_A^*) + \mathcal{L}_{APF}(T', P', T^*, P^*). \tag{3}$$

Since our network is learned with two individual loss functions, the learning schedule and initialization scheme are critical issues. Similar to [12], to unify the three sub-networks seamlessly, we use a training scheme that alternates between fine-tuning for the region proposal network and then fine-tuning for the inference network, while keeping the proposal fixed, enabling a fast convergence and a unified network. As the first stage of the iteration, we set a class label as a random label, and the CWF layer is not learned because it is unstable in the initial learning. During training, the supervision of class variable S^* is determined in an unsupervised manner, which will be described in the following section.

4.3.1. Region proposal network loss

For training the region proposal network, we assign an anchor class label to each proposal candidate. We assign a positive label ($P_A^*(l) = 1$) of the l-th anchor if the anchor is positive, and a negative label ($P_A^*(l) = 0$) if the anchor is negative. With these definitions, we minimize an objective function of the following multi-task (classification and regression) loss such that

$$\begin{aligned} \mathcal{L}_{RPN}(T, P_A, S, T^*, P_A^*) = \mathcal{L}_{RPN}^{CWF}(S) \\ + \lambda_{RPN}^{cls} \sum\nolimits_l \mathcal{L}_{RPN}^{cls}(P_A(l), P_A^*(l)) \\ + \lambda_{RPN}^{reg} P_A^*(l) \sum\nolimits_l \mathcal{L}_{RPN}^{reg}(T(l), T^*(l)), \end{aligned} \tag{4}$$

where λ_{RPN}^{cls} and λ_{RPN}^{reg} are balancing parameters between classification and regression loss, respectively. The classification loss $\mathcal{L}_{RPN}^{cls}$ is the log loss over kinds of anchors. For the regression loss, we use $\mathcal{L}_{RPN}^{reg}(T(l), T^*(l)) = \Omega(R(l) - R^*(l))$ where Ω is the robust loss function (smooth L_1) defined in [11]. The regression loss is activated only for positive sample ($P_A^*(l) = 1$); otherwise, it is disabled ($P_A^*(l) = 0$). For bounding box regression, we adopt the parameterization of transformation vector of the four coordinates $R(l) = [r_{\mathbf{x}}, r_{\mathbf{y}}, \nabla r_{\mathbf{x}}, \nabla r_{\mathbf{y}}]^T$ following [12], defined such that

$$\begin{aligned} r_{\mathbf{x}} &= (l_{\mathbf{x}} - l_{\mathbf{x}}^A)/\nabla l_{\mathbf{x}}^A, \quad r_{\mathbf{y}} = (l_{\mathbf{y}} - l_{\mathbf{y}}^A)/\nabla l_{\mathbf{y}}^A, \\ \nabla r_{\mathbf{x}} &= \log(\nabla l_{\mathbf{x}}/\nabla l_{\mathbf{x}}^A), \quad \nabla r_{\mathbf{y}} = \log(\nabla l_{\mathbf{y}}/\nabla l_{\mathbf{y}}^A), \end{aligned} \tag{5}$$

where $[l_{\mathbf{x}}^A, l_{\mathbf{y}}^A, \nabla l_{\mathbf{x}}^A, \nabla l_{\mathbf{y}}^A]^T$ represents the l-th anchor box. Similarly, for a ground truth box $T^*(l) = [l_{\mathbf{x}}^*, l_{\mathbf{y}}^*, \nabla l_{\mathbf{x}}^*, \nabla l_{\mathbf{y}}^*]^T$, the transformation vector $R^*(l)$ also can be computed.

Furthermore, to select an optimal feature channel among three-branch feature extraction networks at each proposal, the CWF layer employs the additional classification loss $\mathcal{L}_{RPN}^{CWF}$ as a log loss function. However, unlike other loss functions, the ground truth label S^* cannot be estimated. To address this limitation, we formulate a weakly-supervised learning scheme, where tentative label S^* is determined during the iteration to produce the highest probability. By using this tentative class label S^*, we learn the CWF layer in a weakly-supervised manner.

4.3.2. Accumulated probability fusion loss

Similar to the region proposal network loss $\mathcal{L}_{RPN}$, the APF layer loss $\mathcal{L}_{APF}$ consists of classification loss and regression loss functions. However, unlike $\mathcal{L}_{RPN}$, $\mathcal{L}_{APF}$ is defined with accumulated probability, which produces more reliable pedestrian probability. Concretely, we also minimize the following the multi-task loss such that

$$\begin{aligned} \mathcal{L}_{APF}(T', P', T^*, P^*) = \lambda_{APF}^{cls} \mathcal{L}_{APF}^{cls}(P', P^*) \\ + \lambda_{APF}^{reg} P^* \mathcal{L}_{APF}^{reg}(T', T^*), \end{aligned} \tag{6}$$

Table 1
Network configuration of our multi-spectral pedestrian pedestrian method.

	Color/thermal feature extraction network												
	Conv1-1	Conv1-2	Conv2-1	Conv2-2	Conv3-1	Conv3-2	Conv3-3	Conv4-1	Conv4-2	Conv4-3	Conv5-1	Conv5-2	Conv5-3
kernel	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3	3 × 3
channel	64	64	128	128	256	256	256	512	512	512	512	512	512
stride	1	2	1	2	1	1	2	1	1	2	1	1	1
	Fusion feature extraction network				Region proposal network				Inference network				
	NIN[a]	Conv5-1	Conv5-2	Conv5-3	sliding	cls	reg	CWF	FC6	FC7	cls	reg	APF
kernel	1 × 1	3 × 3	3 × 3	3 × 3	3 × 3	1 × 1	1 × 1	1 × 1	7 × 7	1 × 1	1 × 1	1 × 1	1 × 1
channel	512	512	512	512	512	18	36	3	1024	4096	2	8	2
stride	1	1	1	1	1	1	1	1	-	-	-	-	-

[a] Network-in-Network layer [18].

where λ_{APF}^{cls} and λ_{APF}^{reg} are balancing parameters between classification and regression loss. $\mathcal{L}_{APF}^{cls}$ and $\mathcal{L}_{APF}^{reg}$ are defined similar to $\mathcal{L}_{RPN}^{cls}$ and $\mathcal{L}_{RPN}^{reg}$ but different to detect the pedestrian bounding box. We assign a positive label ($P^* = 1$) to the proposal candidate that has an intersection-over-union (IoU) overlap higher than 0.7 with ground-truth bounding boxes and a negative label ($P^* = 0$) whose IoU is lower than 0.3. $\mathcal{L}_{APF}^{cls}$ is a log loss employed to classify the pedestrian or non-pedestrian. For the regression loss, we use a parameterization $B = [b_{\mathbf{x}}, b_{\mathbf{y}}, \nabla b_{\mathbf{x}}, \nabla b_{\mathbf{y}}]^T$ similar to that of RPN such that $\mathcal{L}_{APF}^{reg}(T', T^*) = \Omega(B - B^*)$, where B and B^* are computed similar to (5) with the standard of the pedestrian proposal T. By simultaneously minimizing the overall loss functions of $\mathcal{L}_{RPN}$ and $\mathcal{L}_{APF}$, our overall network is learned to provide optimal pedestrian detection performance.

5. Experimental results and discussion

5.1. Experimental settings

For our experiments, our network was implemented using the TensorFlow machine learning library [32], and it was trained on a NVIDIA GeForce GTX TITAN X GPU. For three-branch feature extraction networks, we used the ImageNet pre-trained VGG-Net [31] from the bottom 'Conv1' to the 'Conv5-3' layer as initial parameters. In our experiments, our network was implemented with the following fixed parameter settings for all datasets: $\{H, W\} = \{7, 7\}$, $\{\lambda_{RPN}^{cls}, \lambda_{RPN}^{reg}, \lambda_{APF}^{cls}, \lambda_{APF}^{reg}\} = \{1, 10, 1, 10\}$, and the number of anchors was set to 9 similar to [12,18]. Furthermore, the sizes of the anchors were defined as 32, 64, and 128, and the ratios of the anchors were set to 0.5, 1, and 2. The number of neighboring proposals was set to 5. To train the network, the standard stochastic gradient descent with momentum was employed for optimization, where the initial learning rate, momentum, and weight decay were set to 0.001, 0.9, and 0.0005, respectively. By default, an IoU threshold of 0.5 was used for determining "true positives". The NMS was applied to the APF layer in order to avoid redundant detections.

5.2. Analysis of probabilistic fusion techniques

We first evaluated two key components in our method, CWF and APF, in comparison to conventional fusion strategies, as presented in Table 2. For quantitative evaluations, we used the KAIST multi-spectral pedestrian dataset [20] for a quantitative evaluation with the miss rate, which will be described in details in the following section. The fusion based approach (i.e., half-way fusion [18]) has shown more stable performances, when compared to the existing detection method [12] in a color or thermal image only. Unlike these approaches, our proposed method can be interpreted to consider all these three networks, i.e., two independent networks for each modality and one fused network. We first evaluated the CWF layer (without the APF layer) in our method in comparison to direct fusion strategies, such as average, multiplication, and max fusion. As expected, these simple fusion schemes cannot provide stable performances for all scenarios; however, they perform rather worse than existing fusion techniques [18]. Unlike these, our CWF method has shown outstanding performances for all scenarios such as daytime and nighttime, reducing the miss rate by 4.35%. Even though the integration of A_I, A_F, and A_C reduced the miss rate of CWF slightly, our CWF that uses only A_C is still employed due to its efficiency. Furthermore, this detection performance can be boosted by considering the neighboring pedestrian proposals. Compared to BBV [23], our APF concentrates on suppressing false positives, thus improving the detection performance dramatically, and this reduces the miss rate of the CWF method by 1.25%.

Table 2
Evaluation of our proposed method by varying the fusion strategies on the KAIST multi-spectral benchmarks [20]. We measure an average miss rate (%) for qualitative evaluations.

Methods	Daytime	Nighttime	All-day
Ren et al. [12] on *I*	43.17	71.40	51.86
Ren et al. [12] on *F*	46.76	35.57	43.04
Liu et al. [18] on *I, F*	38.14	34.42	36.96
Average fusion	36.42	37.21	36.43
Multiplication fusion	37.80	38.90	37.95
Max fusion	35.83	36.01	35.65
CWF	32.87	32.52	32.61
CWF with A_I, A_F, A_C	32.83	32.45	32.56
CWF+BBV [23]	32.63	32.37	32.45
CWF+APF	**31.79**	**30.82**	**31.36**

5.3. Analysis of fusion feature representation

In order to evaluate the performance gain when combining various levels of features and fusion schemes, we additionally compared the multi-spectral fusion methods by varying the level of convolutional features on the KAIST multi-spectral pedestrian dataset [20] in Table 3. For these experiments, we evaluated the results using a single-level feature of 'Conv5' and a multi-level feature of concatenated 'Conv4/5.' Furthermore, we compared various multi-spectral fusion schemes such as simple concatenation, half-way fusion [18], and our CWF+APF.

Through these analyses, we derived two observations for the fusion feature representation. First, the detection accuracy can be improved when using a multi-level feature ('Conv4/5') within all fusion schemes, exemplified in [33,34]. However, the usage of a multi-level feature also requires higher computational complexity compared to that of a single-level feature. Thus, although the usage of multi-level feature also provided the best performance with

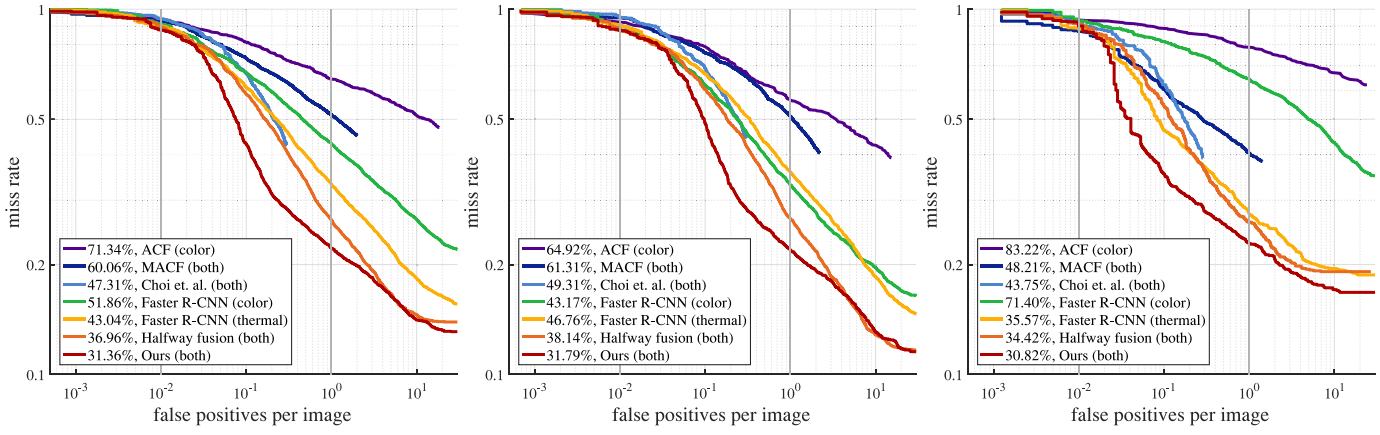
Table 3
Comparison of detection results by varying the feature representation on the KAIST multi-spectral benchmark [20].

Fusion	–	–	Concat.	Half-way [18]	CWF + APF		
Feature	A_I	A_F	A_I, A_F	A_I, A_F	A_I, A_F	$A_I, A_F, A_C^{\dagger}$	A_I, A_F, A_C
Conv5	51.86	43.40	37.33	36.96	34.90	31.54	31.36*
Conv4/5	50.83	42.16	36.75	36.19	34.78	31.43	31.28

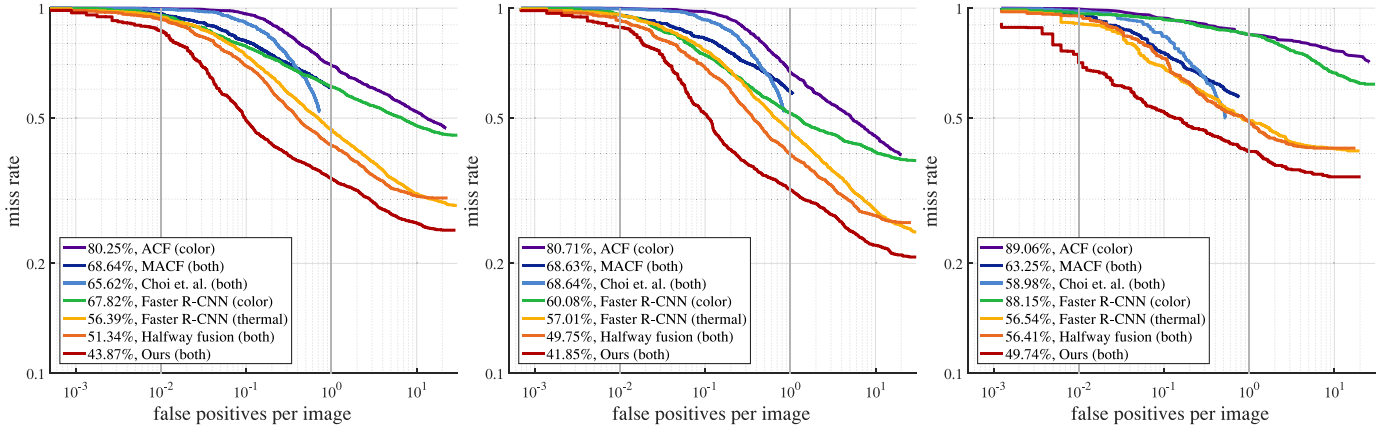
†fusion in 'Conv5', *proposed method

Table 4
Comparison of computation time for handling an image size 640×480.

Methods	Choi et al. [19]	Faster R-CNN [12]	Half-way fusion [18]	Ours
Time (s.)	2.73	0.24	0.43	**0.58**



(a) All-day, all scale (b) Daytime, all scale (c) Nighttime, all scale



(d) All-day, far scale (e) Daytime, far scale (f) Nighttime, far scale

Fig. 7. Comparison of detection results on the test set of KAIST multi-spectral benchmark [20], in terms of (a) all-day, all scale, (b) daytime, all scale, (c) nighttime, all scale, (d) all-day, far scale, (e) daytime, far scale, and (f) nighttime, far scale. Note that in results on the far scale, a pedestrian appears on a small scale, thus providing more challenging scenarios. Compared to other methods, our network provides consistently stable performances across all scenarios.

a small margin in our method, we used a single-level feature of 'Conv5,' considering the trade-off between efficiency and accuracy. Second, with the additional fusion channel (A_C), our CWF+APF approach has shown dramatically improved performances in both cases of using single-level and multi-level feature-based fusions, compared to existing fusion methods such as simple concatenation fusion and half-way fusion [18]. Especially, our method reduced the miss rate by 5.60%, of which 3.54% is reduced by incorporating the fused feature information. Interestingly, when the fusion between multi-spectral channels is established in a deeper convolutional feature of 'Conv5' (denoted $A_C^{\dagger}$), our approach has shown reduced accuracy because the number of layers used to fuse the multi-spectral channels was reduced. Thus, our proposed method is formulated to fuse the multi-spectral channels in 'Conv4' fea-

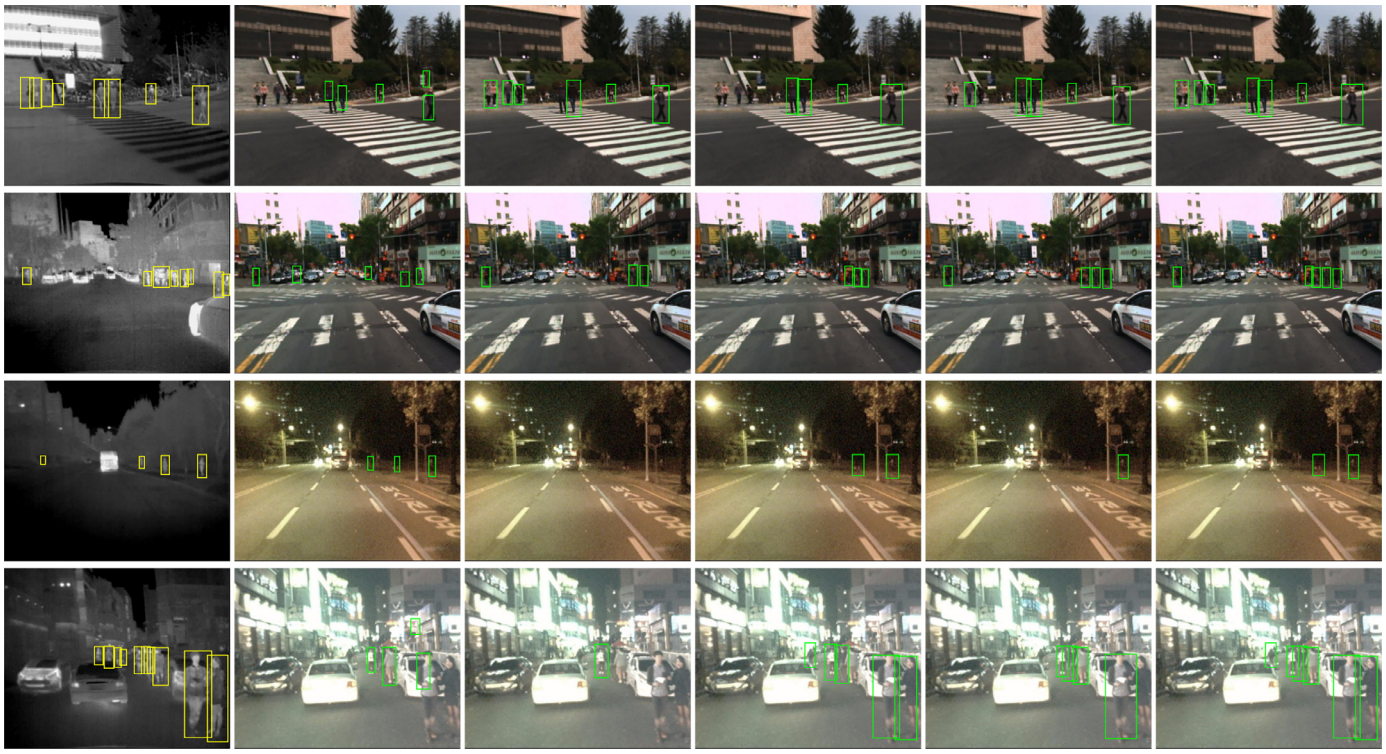


Fig. 8. Qualitative comparisons of detection results on the test set of KAIST multi-spectral dataset [20]. (From left to right) Thermal images with ground-truth, detection results from Choi et al. [19], Faster R-CNN [12] with color and thermal channels, half-way fusion [18], and the proposed method.

ture, followed by additional convolutional layers to make A_C have the same size of A_I and A_F.

5.4. Analysis of computational speed

Table 4 evaluated the computational complexity of our method compared to the state-of-the-art algorithms in handling a color–thermal image pair of a resolution 640×480. Even though our algorithm needs more computational time compared to other previous algorithms, such as Faster R-CNN [12] and half-way fusion [18], it provides the state-of-the-art performance under various challenging situations.

5.5. Evaluation on the KAIST multi-spectral benchmark

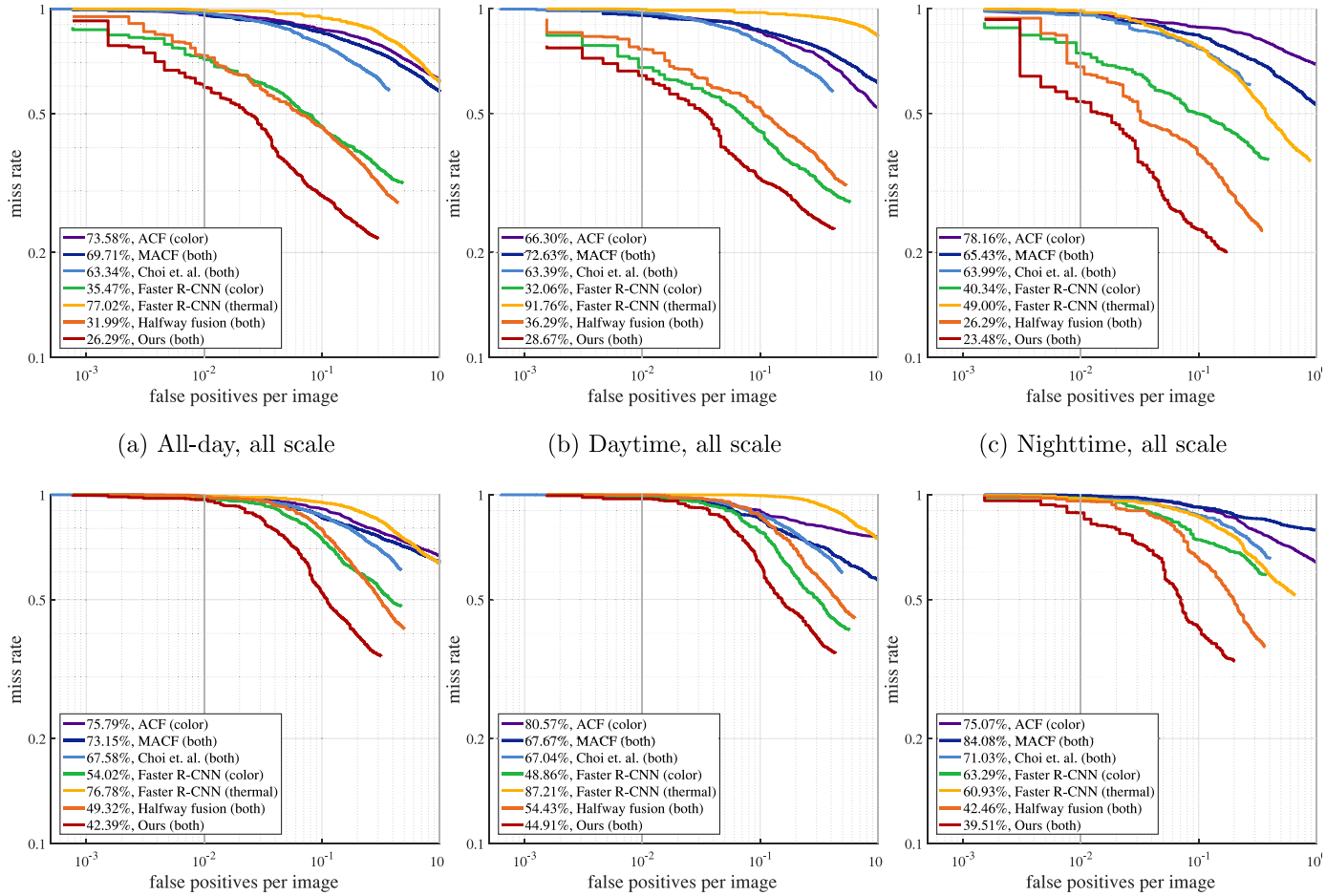
Dataset We evaluated our network on the popular KAIST multi-spectral pedestrian dataset [20], which contains 95,328 aligned color–thermal image pairs with a resolution of 640×480 and 103,128 annotations of pedestrian and cyclist classes. The dataset was taken under various challenging scenarios such as nighttime. The dataset also includes pedestrians in a far scale, which often appears in real-world driving scenarios. For the experiments, we sampled images from training videos with two-frame skips, and obtained 7095 training images. The testing set contains 2252 images sampled from test videos with 30-frame skips, among which 1455 images were captured during daytime and 797 images during nighttime.

Evaluation Figs. 7 and 8 show the quantitative and qualitative comparison results on the KAIST multi-spectral benchmark [20], respectively. For qualitative evaluation, we used a miss rate similar to [20], where a lower miss rate indicates better detection performance on the same false positive per image (FPPI). The results indicate that Faster R-CNN [12] learned in color and thermal channels have shown reasonable performances during daytime, and their fusion scheme [18] also performs well on average. However, these results have shown complementary information for each modality. Thus, we can argue that the fusion of these three-branch fusion methods achieves better performance. Even though Choi et al. [19] has provided reliable performances with its fusion approaches, it also has shown limited performances. Overall, the detection rate was reduced in far scale pedestrians. Unlike these methods, in the quantitative results in Fig. 7, our method achieves the miss rate of 31.36% at all-day detection with all scales, which is considerably better than the best available competitor's 36.96%, which demonstrates that the proposed fusion framework definitely improves the pedestrian detection accuracy.

5.6. Evaluation on the CVC-14 multi-spectral benchmark

Dataset We also evaluated our network on the CVC-14 color–thermal day-night pedestrian sequence dataset [21], which contains 8518 aligned color–thermal sequential image pairs with 9319 pedestrian annotations. This dataset was recorded by on-board color and thermal cameras at 20 Hz, both at a resolution of 640×480. It covered various challenging situations, e.g., many pedestrians at the same time in an image. The testing set of the CVC-14 dataset [21] contains 1433 images, among which 706 images were captured during daytime and 727 during nighttime.

Evaluation Figs. 9 and 10 show the quantitative and qualitative evaluations on the CVC-14 multi-spectral benchmark [21], respectively. Similar to the above experiments, for quantitative evaluations, we measured the miss rate. Our method achieves the miss rate of 26.29% at all-day detection with all scales, which is considerably better than the best available competitor's 31.99%. Interestingly, we observed that the overall miss rates decrease than those in the KAIST benchmark [20]. This is because the CVC-14 benchmark [21] has a higher image quality and a larger pedestrian scale compared to that of the KAIST benchmark [20]. As shown in Fig. 9, our proposed model definitely outperforms other algorithms. Especially, this strength can be shown in the results of the far scale



(a) All-day, all scale (b) Daytime, all scale (c) Nighttime, all scale

(d) All-day, far scale (e) Daytime, far scale (f) Nighttime, far scale

Fig. 9. Comparison of detection results reported on the CVC-14 multi-spectral pedestrian dataset [21], in terms of (a) all-day, all scale, (b) daytime, all scale, (c) nighttime, all scale, (d) all-day, far scale, (e) daytime, far scale, and (f) nighttime, far scale.

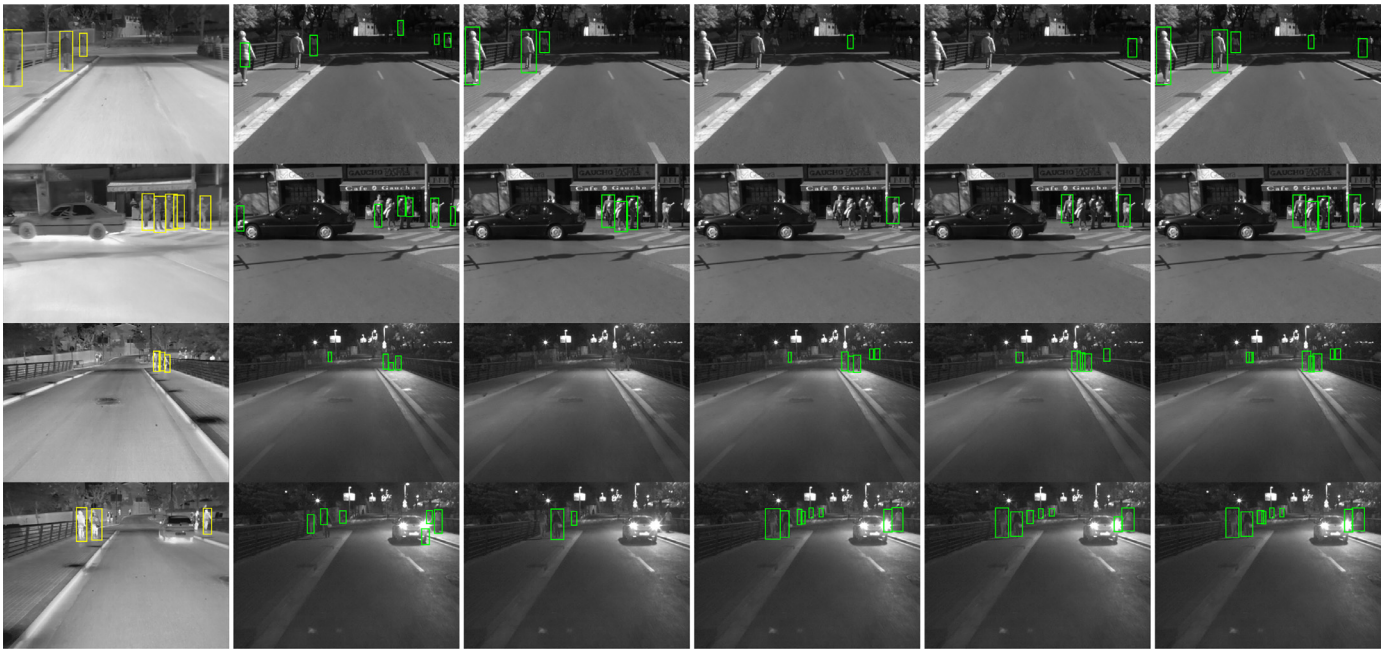
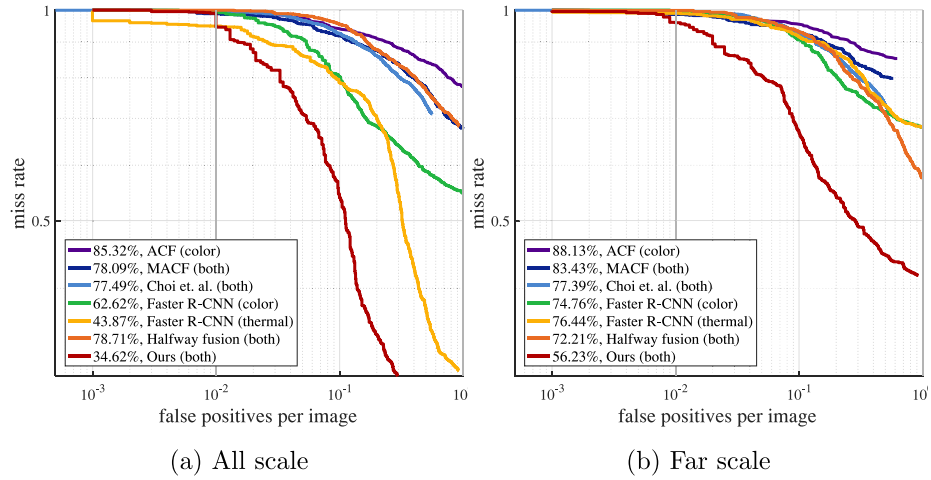


Fig. 10. Examples of detection results on the test set of CVC-14 multi-spectral pedestrian dataset [21]. (From left to right) Thermal images with ground-truth, detection results from Choi et al. [19], Faster R-CNN [12] with color and thermal channels, half-way fusion [18], and the proposed method.



(a) All scale (b) Far scale

Fig. 11. Comparison of detection results reported on the test set of DIML multi-spectral dataset, in terms of (a) all scale and (b) far scale.

Fig. 12. Examples of detection results on the test set of DIML multi-spectral benchmark. (From left to right) The thermal images with ground truth, detection results from [19], Faster R-CNN [12] with color and thermal channels, half-way fusion [18], and the proposed method.

pedestrian cases, which shows that our model also detects a small pedestrian well.

Specifically, through the qualitative comparisons in Fig. 10, we evaluated other characteristics of our fusion method. In the first row of Fig. 10, we can observe that the performance in the color channel only decreases because of the shadow of the surrounding structure even in daytime circumstances. On the other hand, the second row shows that the performance in the thermal channel only was degraded by the surrounding complex environment. These two situations support the necessity of color–thermal fusion. The third row shows that a very small pedestrian, which cannot be shown in the KAIST benchmark [20], can be detected using our method. The last row represents a challenging situation that can occur under actual driving situations, where the detection on the color channel frequently fails. To summarize, the results of the CVC-14 dataset [21] demonstrated that the our fusion method works satisfactorily under challenging driving conditions.

5.7. Evaluation on the DIML multi-spectral benchmark

Dataset We finally evaluated our network on our multi-spectral pedestrian detection benchmark built for indoor surveillance scenarios. In particular, the DIML multi-spectral dataset contains 1003 aligned color–thermal sequential frame pairs with 1792 pedestrian annotations. This dataset was recorded by color and thermal cameras at 10 Hz, both at a resolution of 640×480. We used an uNova20C and an FLIR A65 camera.

Evaluation For these experiments, we used the network model pre-trained on the KAIST benchmark [20], which has similar characteristics to our benchmark. In quantitative evaluations of Fig. 11, our method achieves a miss rate of 34.62% at all scales, considerably better than the best available competitor's 43.87%, thereby showing that our algorithm works satisfactorily even in an indoor surveillance system. Fig. 12 shows the qualitative evaluations, and enables us to evaluate the pedestrian detection performance of

methods in challenging situations of the surveillance system, such as ambient temperature and light saturation.

6. Conclusion

We presented the unified convolutional neural network architecture for multi-spectral color and thermal pedestrian detection even under challenging environments such as nighttime. In contrast to previous techniques, we adopted three-branch detection models taking different image modalities as inputs. To fuse this information simultaneously in a boosting manner, we proposed channel weighting fusion layer and accumulated probability fusion layer, formulated these sub-networks into a single network, and trained the whole network in an end-to-end manner. Our extensive evaluations demonstrated that the proposed method outperforms the state-of-the-arts on challenging multi-spectral pedestrian datasets. We believe that our proposed model can potentially benefit other multi-spectral computer vision tasks in autonomous driving systems or surveillance systems.

Acknowledgment

This work was supported by Institute for Information &Communications Technology Promotion(IITP) grant funded by the Korea government(MSIP) (R0124-16-0002, Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly).

References

[1] M. Enzweiler, D.M. Gavrila, Monocular pedestrian detection: survey and experiments, IEEE Trans. Pattern Anal. Mach. Intell. 31 (12) (2009) 2179–2195.
[2] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.
[3] P. Dollár, R. Appel, S. Belongie, P. Perona, Fast feature pyramids for object detection, IEEE Trans. Pattern Anal. Mach. Intell. 36 (8) (2014) 1532–1545.
[4] P. Dollár, Z. Tu, P. Perona, S. Belongie, Integral channel features, in: Proceedings of the British Machine Vision Conference, 2009.
[5] H. Chen, N. Zheng, J. Qin, Pedestrian detection using sparse Gabor filter and support vector machine, in: Proceedings of the IEEE Intelligent Vehicles Symposium, 2005.
[6] J. Marin, D. Vazquez, A.M. Lopez, J. Amores, B. Leibe, Random forests of local experts for pedestrian detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2013.
[7] D. Ribeiro, J.C. Nascimento, A. Bernardino, G. Carneiro, Improving the performance of pedestrian detectors using convolutional learning, Pattern Recognit. 61 (2017) 641–649.
[8] J. Zhu, O. Javed, J. Liu, Q. Yu, H. Cheng, H. Sawhney, Pedestrian detection in low-resolution imagery by learning multi-scale intrinsic motion structures (MIMS), in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[9] R. Girshick, J. Donahue, T. Darrell, J. Malik, Region-based convolutional networks for accurate object detection and segmentation, IEEE Trans. Pattern Anal. Mach. Intell. 38 (1) (2015) 142–158.
[10] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 37 (9) (2015) 1904–1916.
[11] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015.
[12] 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. 39 (6) (2017) 1137–1149.
[13] Z. Cai, Q. Fan, R.S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: Proceedings of the European Conference on Computer Vision, 2016.
[14] S.G. Kong, J. Heo, F. Boughorbel, Y. Zheng, B.R. Abidi, A. Koschan, M. Yi, M.A. Abidi, Multiscale fusion of visible and thermal ir images for illumination-invariant face recognition, Int. J. Comput. Vis. 71 (2) (2007) 215–233.
[15] J. Han, B. Bhanu, Human activity recognition in thermal infrared imagery, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop, 2005.
[16] J. Han, B. Bhanu, Fusion of color and infrared video for moving human detection, Pattern Recognit. 40 (6) (2007) 1771–1784.
[17] 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.
[18] J. Liu, S. Zhang, S. Wang, D.N. Metaxas, Multispectral deep neural networks for pedestrian detection, in: Proceedings of the British Machine Vision Conference, 2016.
[19] H. Choi, S. Kim, K. Park, K. Sohn, Multi-spectral pedestrian detection based on accumulated object proposal with fully convolution network, in: Proceedings of the IEEE International Conference on Pattern Recognition, 2016.
[20] S. Hwang, J. Park, N. Kim, Y. Choi, I.S. Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[21] A. González, Z. Fang, Y. Socarras, J. Serrat, D. Vzquez, J. Xu, A.M. López, Pedestrian detection at day/night time with visible and fir cameras: a comparison, Sensors 16 (6) (2016) 1–11.
[22] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012.
[23] S. Gidaris, N. Komodakis, Object detection via a multi-region and semantic segmentation-aware CNN model, in: Proceedings of the IEEE International Conference on Computer Vision, 2015.
[24] P. Sermanet, K. Kavukcuoglu, S. Chintala, Y. Lecun, Pedestrian detection with unsupervised multi-stage feature learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[25] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, S. Yan, Scale-aware fast R-CNN for pedestrian detection, in: arXiv:1510.08160, 2015.
[26] Y. Tian, P. Luo, X. Wang, X. Tang, Pedestrian detection aided by deep learning semantic tasks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[27] A. Torabi, G.A. Bilodeau, Local self-similarity-based registration of human ROIs in pairs of stereo thermal-visible videos, Pattern Recognit. 46 (2) (2013) 578–589.
[28] S. Kim, D. Min, B. Ham, M.N. Do, K. Sohn, DASC: robust dense descriptor for multi-modal and multi-spectral correspondence estimation, IEEE Trans Pattern Anal. Mach. Intell. 39 (9) (2017) 1712–1729.
[29] C. Feng, S. Zhuo, X. Zhang, L. Shen, S. Süsstrunk, Near-infrared guided color image dehazing, in: Proceedings of the IEEE International Conference on Image Processing, 2013.
[30] B. Ham, M. Cho, J. Ponce, Robust guided image filtering using nonconvex potentials, IEEE Trans. Pattern Anal. Mach. Intell. 40 (1) (2017) 192–207.
[31] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: arXiv:1409.1556, 2014.
[32] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, et al., Tensorflow: large-scale machine learning on heterogeneous distributed systems, 2016, Software available from http://tensorflow.org/.
[33] S. Bell, C.L. Zitnick, K. Bala1, R. Girshick, Inside-outside net: Detecting objects in context with skip pooling and recurrent neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.
[34] T. Kong, A. Yao, Y. Chen, F. Sun, Hypernet: towards accurate region proposal generation and joint object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

Kihong Park received the B.S. degree in electronic engineering from Sogang University, Seoul, Korea, in 2014. He is currently pursuing the joint M.S. and Ph.D. degrees in electrical and electronic engineering at Yonsei University. His current research interests include computer vision, and machine learning, in particular, depth estimation, multi-modal detection.

Seungryong Kim received the B.S. and Ph.D. degrees in Electrical and Electronic Engineering from Yonsei University, Seoul, Korea, in 2012 and 2018, respectively. He is currently a Post-Doctoral Researcher in Electrical and Electronic Engineering at Yonsei University. His current research interests include 2D/3D computer vision, computational photography, and machine learning.

Kwanghoon Sohn received the B.E. degree in electronics engineering from Yonsei University, Seoul, Korea, in 1983, the M.S.E.E. degree in electrical engineering from University of Minnesota in 1985, and the Ph.D. degree in electrical and computer engineering from North Carolina State University in 1992. He is currently a professor in the School of Electrical and Electronic Engineering at Yonsei University.