LOW-LIGHT PEDESTRIAN DETECTION FROM RGB IMAGES USING MULTI-MODAL KNOWLEDGE DISTILLATION

Srinivas S S Kruthiventi, Pratyush Sahay and Rajesh Biswal

Harman International Industries, India

srinivas.sai@harman.com, pratyush.sahay@harman.com, rajesh.biswal@harman.com

ABSTRACT

While deep learning based pedestrian detection systems have continued to scale new heights in recent times, the performance of such algorithms tends to degrade under challenging illumination conditions. This causes a bottleneck in ready portability of such systems to Advanced Driver Assistance Systems (ADAS), where consistent performance across varying environmental lighting is desired. Inspired by the concept of *dark knowledge*, this paper proposes a novel guided deep network that distills knowledge from a multi-modal pedestrian detector. The proposed network learns to extract both *RGB and thermal-like* features from RGB images alone, thus compensating for the requirement of significantly costly automotive-grade thermal cameras. Compelling detection performance in severe lighting conditions is demonstrated on a publicly available night-time pedestrian dataset - KAIST. We achieve an effective miss-rate of 12% lower than the recent state-of-the-art methods.

Index Terms—
Convolutional Neural Networks, Low-light pedestrian detection, Knowledge Distillation

1. INTRODUCTION

The importance of consistently performing pedestrian detection algorithms across varying lighting conditions cannot be overstated. Such systems form the core components of applications such as automated driving, ADAS, visual surveillance etc. However, top algorithms on Caltech benchmark [1], such as computer vision based models [2, 3] and deep learning based models [4, 5], tend to underachieve when tasked with pedestrian detection under challenging illumination conditions. Multi-modal techniques [6, 7, 8] incorporating additional data from a thermal camera, depth cameras, etc. have been proposed in various vision-based tasks to overcome the ill-effects of ambient lighting on RGB data. However, reliance on extra hardware tends to increase the setup costs for the system significantly, especially so in the case of automotive-grade solutions.

Knowledge distillation [9] has been successfully demonstrated as a means of transferring the hidden understanding of a cumbersome training model to a simpler deployment-phase model.Shen et al. [10] extend this idea to train a smaller student network for pedestrian detection by utilizing the intermediate hint features of a teacher network and a dual loss function based on the ground-truth labels and these hint features.

Liu et al. [7], in their work of multi-spectral pedestrian detection, established the complementary visual nature of RGB and thermal channels under varying illumination conditions. They developed a Faster R-CNN [11] based halfway fusion model that taps features that are 'best of both the worlds' to perform competitive pedestrian detection even under severe illumination conditions.

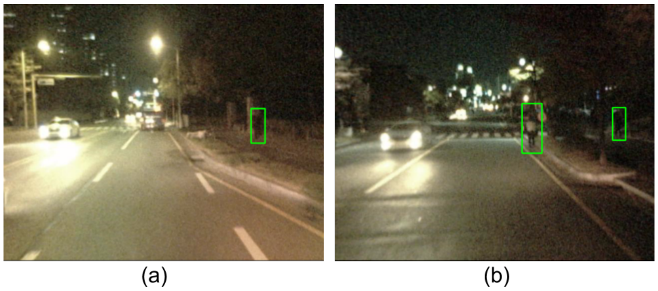

Fig. 1. Sample pedestrian detection performance of the proposed method under challenging illumination conditions. Pedestrians in both (a) low visibility, and (b) medium or high visibility are detected.

In this paper, a novel guided deep network is proposed that learns to extract *multi-modal like* features from a single data modality in the framework of knowledge distillation by means of teacher-student training. A ResNet-50 [12] based deep network is developed that takes in RGB image data, extracts both *RGB and thermal-like* features to effectively detect pedestrians across challenging illumination conditions. The effectiveness of the proposed algorithm is demonstrated on a publicly available night time dataset - KAIST [13], with an effective miss-rate of 12% lower than other recent methods.

Following are the major contributions of our work:
(i) A novel method of extracting *multi-modal like* features from single data modality in the framework of teacher-student network
(ii) A pedestrian detection model which achieves significantly lower miss rates in challenging illumination conditions

2. RELATED WORK

Extraction of complementary modal information from inherent features inferred from RGB data is not new to the vision community, with works such as [14, 15] attempting scene depth recovery from a single RGB image. Walker et al.[16] extracted dense optical flow information from a single static image, while a DNN-based image caption generation model [17] was shown to infer image textual descriptions given an RGB image.

Pedestrian detection is a well-studied topic mostly addressed as a specific case of object detection. Models using both traditional computer vision-based techniques [18, 2, 19, 3] and DNN-based techniques [4, 5] perform well on the Caltech benchmark [1]. Since Caltech dataset consists of video sequences captured under good environmental lighting conditions, the above algorithms tend to under-

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

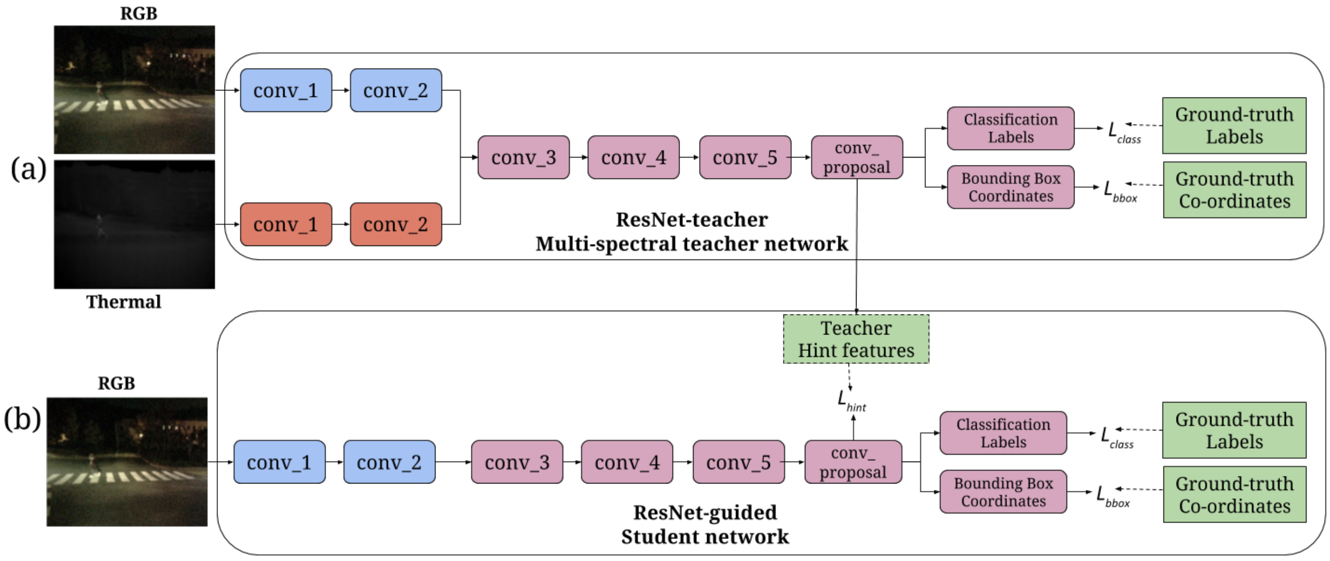

Fig. 2. Overview of the proposed approach using (a) teacher network, for the guided training of (b) student network.

achieve when posed with inputs captured under challenging lighting conditions. Multi-spectral DNN based techniques for pedestrian detection [7, 8] explored developing models that result in effective synergy in the fusion of RGB and thermal data for pedestrian detection.

In contrast, the proposed algorithm learns to extract joint multi-modal features (*RGB and thermal-like*) from RGB data alone under the guidance of a teacher network, and uses these features to produce state-of-the-art pedestrian detection performance across variations in lighting conditions.

3. PROPOSED METHOD

In this work, we propose a novel approach for pedestrian detection in challenging illumination conditions using only RGB images in the framework of teacher-student training. The teacher network, trained using multi-modal data of RGB and thermal images, guides the student network to extract *RGB and thermal-like* features from RGB images alone. Features learned by the teacher network, often termed as *dark knowledge*, contain rich information about the visual characteristics of pedestrians that can be extracted from both RGB and thermal images. The student network is trained to leverage upon this dark knowledge by trying to mimic the features of the teacher network while relying only on RGB images.

The core idea behind developing such a network stems from the observation that in regions with low-to-medium visibility, the visual features discriminating pedestrians from the background are feeble, but still present in RGB images. A network trained directly on the RGB images in an unguided manner may miss out on capturing such features. In comparison, a network guided with the dark knowledge of the multi-modal teacher can be trained to extract such discriminative features from a region even if they are only feebly present.

The training of the proposed network proceeds in two phases - (a) the teacher network is trained on both RGB and thermal images, and (b) the student network is trained on RGB images alone while incorporating knowledge distilled from the trained teacher network. It is to be noted that once the teacher network has been trained, its weights are frozen and remain unchanged during the training of the student network. Details about these individual networks and teacher-student training process are described in the following sub-sections.

3.1. Teacher Network

The teacher network, referred to as *ResNet-teacher*, is a multi-modal pedestrian detector as depicted in Fig. 2(a). It uses ResNet-50 as the base model, which is a 50 layers deep incarnation of ResNet classification network [12]. The initial blocks of *ResNet-teacher* are modified to have two separate arms which individually tap features from both RGB and thermal images. The extracted RGB and thermal features are fused after the conv_2 block, and these fused features are propagated across the following blocks of conv_3, conv_4 and conv_5.

Following the conv_5 block, we introduce a 3×3 convolutional layer with 512-channels, termed as conv_proposal layer. Features from this layer, $F_{teacher} \in \mathbb{R}^{w \times h \times 512}$, are used to predict classification labels and bounding-box coordinates at each spatial location in $F_{teacher}$. We achieve this by processing $F_{teacher}$ through two sibling Region Proposal Network (RPN) arms of classification and regression. Given the single category detection problem at hand, i.e., *pedestrian*, the RPN itself suffices as the object detector. Similar to [5], the RPN anchors at each spatial location are of single aspect ratio 0.41, and of 9 different scales.

Since the conv_proposal layer is immediately followed by the classification and regression arms, the output feature from this layer, $F_{teacher}$, serves as a rich feature descriptor assimilating the *dark knowledge* of *ResNet-teacher* needed for pedestrian detection. Given its richness, we use this feature to guide the student network.

3.2. Student Network

The architecture of the student network, referred to as *ResNet-guided*, shown in Fig. 2(b), is similar to *ResNet-teacher* except for the usage of RGB image alone as the input. Similar to $F_{teacher}$, features from the conv_proposal layer of *ResNet-guided*, i.e.,

$F_{student} \in \mathbb{R}^{w \times h \times 512}$, form the essence of this network's learning. In order to efficiently perform pedestrian detection under challenging illumination conditions, $F_{student}$ should be similar to $F_{teacher}$ for a given input image. This is enforced by penalizing the difference between $F_{teacher}$ and $F_{student}$ in the smooth $L1$-loss sense:

$$\mathbb{L}_{hint} = \|F_{teacher} - F_{student}\|_1 \tag{1}$$

This loss in Eq. 1, called the hint loss, tends to equally weigh all the spatial locations in the $w \times h$ feature map. However, such an equal weighting of the loss is counter-productive since regions with zero-to-low visibility hardly contain any useful data from which discriminative features can be extracted. To overcome this, we introduce a weighted $L1$-hint loss, where the weight for each region is calculated based on its visibility. We define the visibility of a region, $\mathcal{V} \in \mathbb{R}^{w \times h}$, as the entropy of its intensity histogram. Higher entropy obtained for regions with greater visual detail implies higher visibility levels, while a lower value of entropy signifies low visibility.

$$\begin{aligned} \mathcal{V}(i,j) &= \mathcal{E}(\text{hist}(I(\mathbf{i}_p, \mathbf{j}_p)) && (2) \\ \text{where : } \mathcal{E}(\mathbf{w}) &= -\mathbf{w}^T \log(\mathbf{w}) && (3) \end{aligned}$$

and hist$(I(\mathbf{i}_p, \mathbf{j}_p)$ is the histogram of a $p \times p$ region in the input image centered at a location corresponding to (i, j) in the feature map $F_{student}$.

The values obtained in the visibility map $\mathcal{V}$ are normalized to lie between $[0, 1]$. Also, to match the channel-size of $F_{student}$, $\mathcal{V}$ is replicated across 512-channels. Hence, the modified hint loss is:

$$\mathbb{L}_{hint} = \|(F_{teacher} - F_{student}) \times \mathcal{V}\|_1 \tag{4}$$

The hint loss as given by Eq. 4 ensures that image regions with higher visibility contribute more to the hint loss, compared to image regions with lower visibility. This enables extraction of *RGB and thermal-like* features from regions which are even reasonably visible in the RGB image.

3.3. Combined loss function for *ResNet-guided*

In addition to the weighted hint loss $\mathbb{L}_{hint}$, in Eq. 4, we also backpropagate gradients based on the loss from ground-truth annotations for classification labels $\mathbb{L}_{class}$ and bounding box coordinates $\mathbb{L}_{bbox}$. Thus, the combined loss $\mathbb{L}$ for the training of the student network *ResNet-guided* is given as:

$$\mathbb{L} = \alpha \mathbb{L}_{bbox} + \beta \mathbb{L}_{class} + \gamma \mathbb{L}_{hint} \tag{5}$$

where, $\alpha, \beta, \gamma \in \mathbb{R}$, are the relative weights assigned to the $\mathbb{L}_{bbox}$, $\mathbb{L}_{class}$ and $\mathbb{L}_{hint}$ respectively.

4. EXPERIMENTAL EVALUATION

We evaluate the proposed approach for low-light pedestrian detection on the test set of publicly available KAIST Multi-spectral Pedestrian dataset [13]. This dataset is collated from the video-feed of a camera mounted on a car. It contains scenarios of driving at night-time in *campus, urban* and *downtown* localities captured in 6 sets of videos. This dataset provides data in both RGB and thermal modalities for each frame in the video along with dense bounding-box annotations of the pedestrians. Of the 6 video sets provided, we use the standard split of 3 sets for training, validation and the rest for testing the performance of the proposed approach. We sample alternate frames from the train video sets resulting in a total of $7,192$ images for training.

As described in Sec. 3, our model is based on the ResNet-50 classification network. This network outputs bounding-box region proposals along with their confidence scores at each anchor for pedestrian detection. Since the model deals with a two-class object detection problem (*pedestrian* vs. *background*), the ResNet-50 based region proposal network alone can be used as a potential object detector.

Before training the model on KAIST dataset, we initialize the ResNet-50 based detection network by training it for day-time detection on approximately 45k images of Caltech Pedestrian dataset [1]. This pre-training is found to provide a good initialization to the network for pedestrian detection in night-time. With this initialization, *ResNet-teacher*, described in Sec. 3.1, is trained using both the RGB and thermal modalities of KAIST night-time dataset. We have reduced the stride of `conv_5` block in the base ResNet-50 to 1 in order to perform detection at an effective stride of 16 on the original image. This removal of stride is compensated using the filter dilation technique of [20]. *ResNet-teacher* is trained with a mini-batch size of 1, learning rate of 5×10^{-4}, weight-decay of 1×10^{-3} and momentum of 0.9. The training is carried out for $10,000$ iterations and the final model is obtained by choosing the best performing model over the validation set. We evaluate the performance of the model using the standard metric of Average Miss Rate [1]. *ResNet-teacher* is observed to give an average miss rate of 29%. This metric takes the average of miss rates across False Positive Per Image (FPPI) rates between 10^{-2} to 10^{0} obtained by thresholding the model's detections at various confidence cut-off values.

ResNet-teacher guides the training of *ResNet-guided* by providing hint features - $F_{teacher}$, as described in Sec. 3. The relative weights for various losses, shown in Eq. 5, are chosen so as to keep the individual losses in a similar range. These are heuristically chosen to be $\alpha = 5, \beta = 1, \gamma = 5 \times 10^{-4}$. The training hyper-parameters for *ResNet-guided* are kept the same as that of *ResNet-teacher*. We would like to re-emphasize that *ResNet-guided* is trained to detect pedestrians only using RGB images and does not require any input from thermal sensors during deployment of the model.

4.1. Results

In Table 1, we present the Average Miss Rate of our student model - *ResNet-guided*, on KAIST night-time test set. We compare the performance of our model with other recent approaches for pedestrian detection in night-time which use only RGB images. Further, we also compare the performance of the proposed approach of guided training against a network trained directly from the ground-truth without any hint features. This model is labeled as *ResNet-direct*. As illustrated in Table 1, our network outperforms the other recent works by a huge margin. We also obtain a significant gain (4% relative improvement) by training the student network in a guided manner as opposed to direct training.

The qualitative results of *ResNet-guided* on KAIST test set are shown in Fig. 3. The green boxes indicate the pedestrians which are correctly detected by the model. As evident from images in first and second rows of Fig. 3, the model is able to accurately detect pedestrians across a wide range of illumination conditions. Yellow boxes indicate false-positives of the model. It is observed that these false-positives mostly arise from trees, poles, sign-boards being falsely detected. Red boxes in the figure represent pedestrians which are missed by the model. Such pedestrians are often found to have either very poor visibility or suffer from camouflaging/occlusion due to their cluttered background.

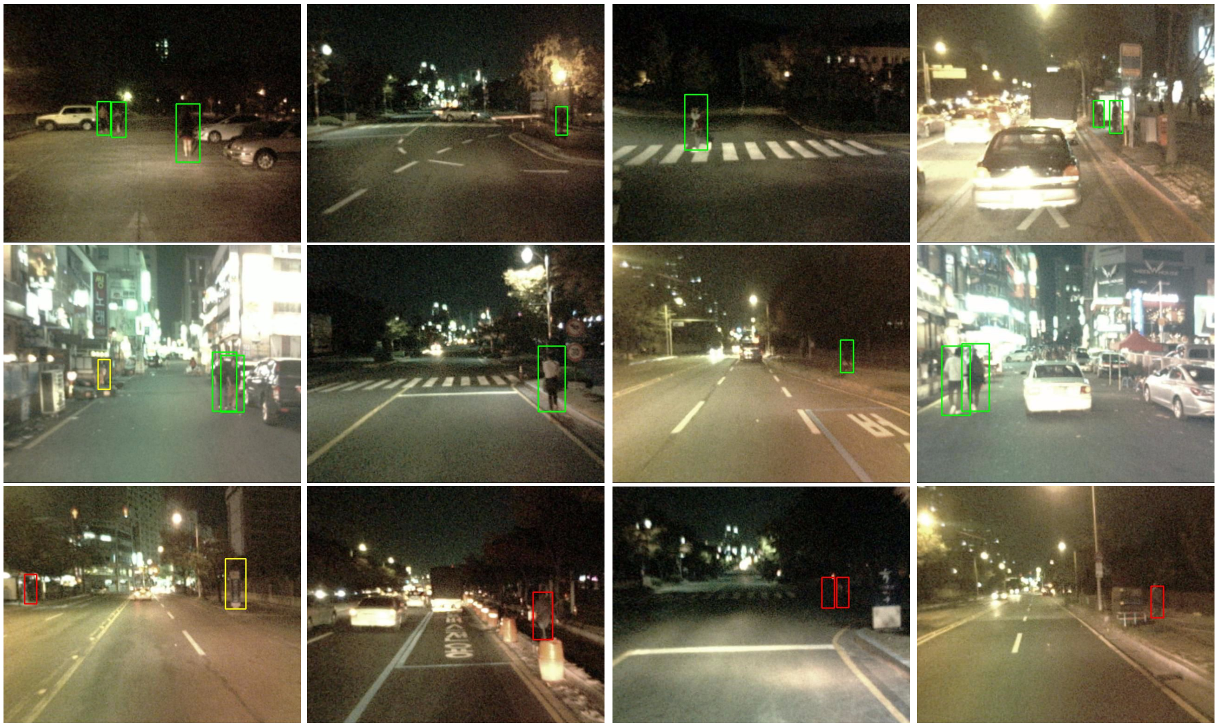

Fig. 3. Qualitative results of the proposed student network – *ResNet-guided*, on test set of KAIST night-time pedestrian dataset. Green and yellow boxes indicate true positives and false positives of the model respectively, while the red boxes indicate pedestrians missed by the model. First two rows illustrate scenarios where the model is doing well at pedestrian detection while the third row shows the failure cases.

Table 1. Quantitative results of our approach in comparison with recent pedestrian detection algorithms on KAIST night-time test set. The algorithms presented perform detection using only RGB images.

Approach	**Average Miss Rate (Lower the better)**
Hwang et al. [13]	90.17%
Wagner et al. [8]	63.40%
Liu et al. [7]	64.39%
ResNet-direct (Ours)	55.26%
***ResNet-guided* (Ours)**	**52.81%**

4.2. Analysis

In order to better understand the performance of the proposed model, we categorize the pedestrians into various bands based on their visibility. As described in Sec. 3.2, we obtain the visibility of a pedestrian patch by calculating the entropy of its histogram. This association of visibility value to the pedestrians allows us to analyze the performance of the model as a function of pedestrian visibility.

In Fig. 4, the bar plot (in green color) illustrates the distribution of pedestrians in the dataset across different bands of visibility ranging from 3.25 to 5.25 in steps of 0.25. In each unnormalized visibility band, we compute the miss rate of the two proposed models – *ResNet-direct* and *ResNet-guided* at a fixed overall FPPI rate of 0.1. From the figure, it can be observed that the performance of both the models improves significantly with an increase in the pedestrian's visibility. As expected, both the models perform equally well in the higher visibility regions. However, in low-visibility regions (< 4.75), *ResNet-guided* can be observed to significantly outperform *ResNet-direct*. This illustrates that *ResNet-guided* effectively leverages the hint features provided by *ResNet-teacher* in detecting pedestrians under challenging illumination conditions.

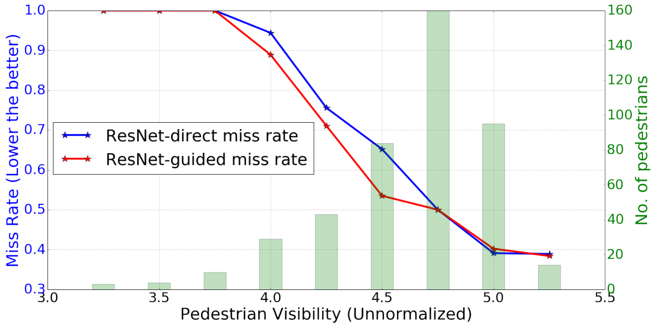

Fig. 4. Miss rate curves of the proposed models (left Y-axis) and pedestrian distribution bars (right Y-axis) in various bands of unnormalized pedestrian visibility. *ResNet-guided* detects pedestrians with low visibility better compared to *ResNet-direct*.

5. CONCLUSION

Our work proposes a deep convolutional network for pedestrian detection under challenging illumination conditions while relying only on RGB images. The proposed model is guided during training by a multi-modal teacher to extract *RGB and thermal-like* features from RGB images alone. Quantitative analysis suggests that the proposed approach of guided training has enabled the model to better detect pedestrians which are only marginally visible. Further, this guided model is experimentally shown to outperform a model trained only from ground-truth annotations and has achieved a significant improvement of 12% over the existing state-of-the-art methods.

6. REFERENCES

[1] Piotr Dollár, Christian Wojek, Bernt Schiele, and Pietro Perona, “Pedestrian detection: An evaluation of the state of the art,” *PAMI*, vol. 34, 2012.

[2] Piotr Dollár, Ron Appel, Serge Belongie, and Pietro Perona, “Fast feature pyramids for object detection,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 8, pp. 1532–1545, 2014.

[3] Rodrigo Benenson, Mohamed Omran, Jan Hosang, and Bernt Schiele, “Ten years of pedestrian detection, what have we learned?,” in *European Conference on Computer Vision*. Springer, 2014, pp. 613–627.

[4] Zhaowei Cai, Quanfu Fan, Rogerio S Feris, and Nuno Vasconcelos, “A unified multi-scale deep convolutional neural network for fast object detection,” in *European Conference on Computer Vision*. Springer, 2016, pp. 354–370.

[5] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He, “Is faster r-cnn doing well for pedestrian detection?,” in *European Conference on Computer Vision*. Springer, 2016, pp. 443–457.

[6] Jun Liu, Ye Liu, Ying Cui, and Yan Qiu Chen, “Real-time human detection and tracking in complex environments using single rgbd camera,” in *Image Processing (ICIP), 2013 20th IEEE International Conference on*. IEEE, 2013, pp. 3088–3092.

[7] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas, “Multispectral deep neural networks for pedestrian detection,” in *British Machine Vision Conference (BMVC)*, 2016.

[8] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke, “Multispectral pedestrian detection using deep fusion convolutional neural networks,” in *European Symp. on Artificial Neural Networks (ESANN)*, 2016.

[9] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean, “Distilling the knowledge in a neural network,” *arXiv preprint arXiv:1503.02531*, 2015.

[10] Jonathan Shen, Noranart Vesdapunt, Vishnu N Boddeti, and Kris M Kitani, “In teacher we trust: Learning compressed models for pedestrian detection,” *arXiv preprint arXiv:1612.00478*, 2016.

[11] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in *Advances in neural information processing systems*, 2015, pp. 91–99.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[13] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon, “Multispectral pedestrian detection: Benchmark dataset and baseline,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1037–1045.

[14] Ashutosh Saxena, Min Sun, and Andrew Y Ng, “Make3d: Learning 3d scene structure from a single still image,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 5, pp. 824–840, 2009.

[15] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem, “Rendering synthetic objects into legacy photographs,” in *ACM Transactions on Graphics (TOG)*. ACM, 2011, vol. 30, p. 157.

[16] Jacob Walker, Abhinav Gupta, and Martial Hebert, “Dense optical flow prediction from a static image,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2443–2451.

[17] Andrej Karpathy and Li Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3128–3137.

[18] Jianfei Dong, Junfeng Ge, and Yupin Luo, “Nighttime pedestrian detection with near infrared using cascaded classifiers,” in *Image Processing, 2007. ICIP 2007. IEEE International Conference on*. IEEE, 2007, vol. 6, pp. VI–185.

[19] Xiaoyu Wang, Tony X Han, and Shuicheng Yan, “An hog-lbp human detector with partial occlusion handling,” in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 32–39.

[20] Chen Liang-Chieh, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan Yuille, “Semantic image segmentation with deep convolutional nets and fully connected crfs,” in *International Conference on Learning Representations*, 2015.