Infrared Pedestrian Detection with Converted Temperature Map

Yifan Zhao*[1], Jingchun Cheng*[2], Wei Zhou[3], Chunxi Zhang[1] and Xiong Pan[1]
1. Beihang University, E-mails: zhaoyifan928, zhangchunxi and 08768@buaa.edu.cn
2. Tsinghua University, E-mail: chengjingchun14@163.com
3. University of Science and Technology of China, E-mail: weizhou_geek@163.com
∗ denotes equal contribution.

***Abstract*—Infrared pedestrian detection aims to detect persons in outdoor thermal images. It shows a unique advantage in dark environment or bad weather compared to daytime visible images (the RGB image). Most current methods treat infrared detection the same way as with visible images, e.g. regarding the infrared image as a special gray-scale visible image. In this paper, we tackle this problem with more emphasis on the underlying temperature information in infrared images. We build an image-temperature transformation formula based upon infrared image formation theory, which can convert infrared image into temperature map with the prior of pedestrian pixel-temperature value. The whole detection process follows a two-stage manner. In the first stage, we use a common detector which treats the infrared image as the gray-scale visible image to provide primary detection results and a pedestrian position prior (the highest-confidence pedestrian detection box in each image). In the second stage, we convert infrared images into corresponding temperature maps and train a temperature net for detection. The final results consist of both the primary detection and the temperature net outputs, detecting pedestrians with characteristics in both image and temperature domain. We show that the converted temperature image is less affected by environmental factors, and that its detector shows amazing complementary ability with the primary detector. We carry out extensive experiments and analysis on two public infrared datasets, the OTCBVS dataset and the FLIR dataset; and demonstrate the effectiveness of incorporating temperature maps.**

Fig. 1. An illustration of differences between infrared and visible images. This figure gives some examples for visible RGB-scale, visible gray-scale, and infrared images taken under the same conditions, i.e. with light pollution, bad weather, severe occlusion, and darkness. We can see that visible images may show more details, while infrared ones are more temperature-sensitive, and more robust to bad conditions.

I. Introduction

Infrared pedestrian detection [1][2][3] has become an arising research hot-spot these years as the infrared images have unique advantages, e.g. specialty in heat feature representation, high robustness to bad conditions, etc. We show some examples of differences between infrared images and visible images in Figure 1, illustrating that infrared images are naturally better at handling occlusion, dark light, and spoofing compared with visible ones. Infrared pedestrian detection is widely used in numerous practical applications, such as autonomous driving [**?**][4][5], video surveillance [6][7][8], night vision [9][10][11], and navigation [12][13][14]. However, it still remains a challenging task, for the real-world infrared images endure all kinds of environmental influences, and can be of low-quality or with cluttered information. In this paper, we propose a novel infrared pedestrian detection model which relieves the impact of external factors by emphasizing on the extracted temperature information.

Numerous methods have been developed in the field of infrared pedestrian detection. They can be broadly categorized into traditional models and deep models. Traditional models [2][3][15] incorporate elaborately-designed hand-crafted features (e.g. Local Binary Patterns, LBP [16] and Histograms of Gradients, HOG [17]) and common classifiers (e.g. SVM and Bayesian classifier) to find and recognize pedestrian areas in infrared images. They usually follow a similar detection pattern: region proposal, region feature extraction, region classification and post-processing. First, the region proposal step generates hundreds of thousands of candidate regions, then the feature extraction step depicts candidate regions with well-designed image descriptions, afterwards a specific classifier classifies them into ”person” or ”non-person”, where the ”person” boxes go through a post-processing method (e.g. the Non-Maximum Suppression algorithm, NMS) and come to the final detection outputs. In contrast to traditional models, deep models [18][19][20][21][22][23][24][25] automatically learn the image high-level abbreviations from large-scale annotated

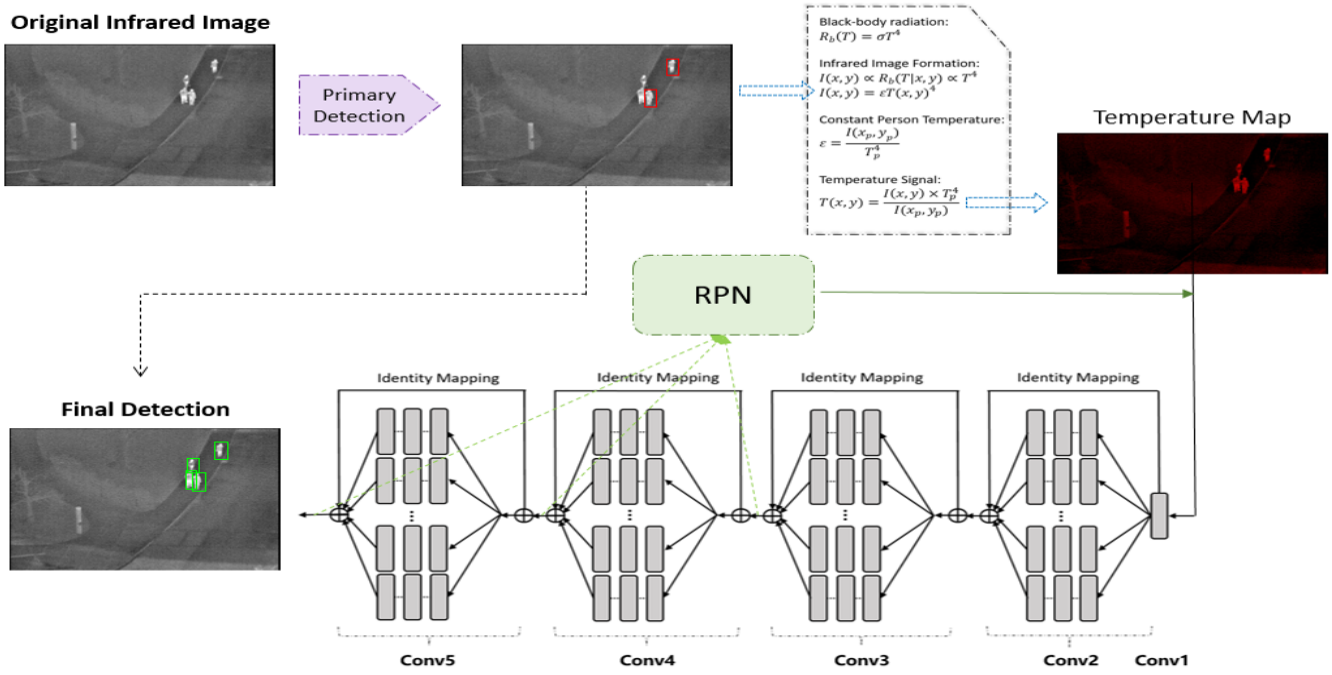

Fig. 2. Overall pedestrian detection framework. This figure shows our overall detection framework. The original infrared image first goes into a primary detector (a detection network trained in the gray-scale visible image pattern); then the image is transformed into the corresponding temperature map according to the primary detection result and the proposed transformation function; after that, a temperature network is trained to detect pedestrians via their temperature characteristics. In the final stage, detection results from both the primary detector and the temperature net are combined as the algorithm output.

datasets. They use convolutional neural networks (CNN)[26] whose parameters are mined from big data to replace the traditional hand-crafted descriptions in the detection procedure, and show overwhelmingly better performance than the traditional models. Moreover, deep models manage to fuse the separate detection steps into a unified framework. For example, the Fast-RCNN [27] detection model utilizes spatial pooling strategy to enable candidate regions to share computation of convolutional layers, speeding up the detection procedure by over a hundred times. And the Faster-RCNN framework [28] takes a step forward by replacing the region proposal step with an RPN (region proposal network) which can be trained simultaneously with the overall detection net. In this work, we use a detection module like Faster-RCNN, where the whole process is modeled into one unified, end-to-end trainable network.

We find that both traditional and deep models treat infrared images like gray-scale visible images. This may introduce erroneous cognition for that the pixel values in infrared images do not contain the same kind of information as visible ones. The difference between infrared images and gray-scale visible images can be seen in Figure 1: the gray-scale images contain detailed human texture information but share the same vulnerability as RGB images, while the infrared images show heat distribution under a certain condition and can intuitively distinguish living beings. Therefore, methods dealing with infrared images the same way as visible ones fail to make full use of their unique thermal information. Based on this observation, we propose that the inherent temperature message should be emphasized more in infrared detection. In this paper, we extract the temperature map from the infrared image, and detect pedestrians upon the clean and normalized temperature image as a supplementary to common detectors. We show that the proposed detection framework is thermally sensitive, and has better robustness and recognition ability than algorithms dealing with infrared images only in the same way as gray-scale visible ones.

Overall, our main contributions are three-fold.

1) We manage to extract the original temperature map from infrared images regardless of external influences.

2) We construct a powerful and robust detection module for general infrared pedestrian detection.

3) We improve performance from a common detector by a large extent with help the our temperature net.

II. Temperature-sensitive Infrared Pedestrian Detection

In this section, we introduce the details of our infrared pedestrian detection model, including the network structure, infrared to temperature transformation formula and the implementation details.

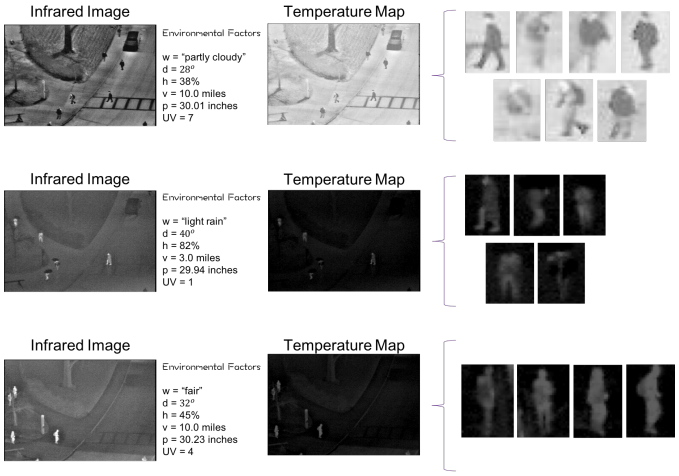

Fig. 3. Illustration of infrared images taken under different environmental conditions. This figure provides three examples of infrared images taken with different environmental factors. w, d, h, v, p, UV denote the environmental factors of weather, dew point, humidity, visibility, pressure and UV index, respectively (as presented in Equation 2). We show the converted temperature maps with pedestrian areas enlarged on the left. Comparing with original infrared images, temperature maps are cleaner and show more pixel value similarities in pedestrian areas. This validates our idea that temperature maps contain more purified and unified thermal information that may help the pedestrian detection.

A. Network Structure

The main detection network follows the structure of Faster R-CNN [28], where candidate regions are proposed by an RPN (Region Proposal Network) which shares the convolutional features with the classification head that categorizes candidate regions and regress to optimal positions. We choose ResNext-101 network [29] as the basic body net for its effectiveness and efficiency in both training and testing (the superiority of ResNext-101 has been validated by a number of works [30][31][32][33][34][35][36][37][38]). This network has five large blocks, each consisting of analogical groups of convolution operations. Figure 2 shows the architecture of our detection network, where rich features for region proposal and classification are extracted from multiple blocks, i.e. Conv3, Conv4, and Conv5. We set the prediction class to be binary: "pedestrian" or "non-pedestrian". During training, the network is optimized with the region proposal loss and classification loss, each consisting of a Cross-Entropy loss for classification and a Smooth-L1 loss for bounding box regression.

B. From Infrared Image to Temperature Map

The key point in the proposed framework is to extract the temperature information from infrared images with discrete pixel values (0, 1, ... 255). The intuitive idea is that pixel values may be proportional to temperature, which, is proved to be otherwise as shown in Figure 3. It clearly illustrates that a certain pixel value in infrared images does not directly correspond to a certain temperature value.

In order to get the relationship between pixel values and temperatures, we go back to the infrared image formation principles. Infrared images are obtained by non-contact devices which detect and convert infrared radiations into electronic signals (quantified to pixel value for display). As introduced in [39], this radiation energy to pixel value transformation is proportional (we verify this on the OTCBVS dataset in Figure 4). The radiation energy that reaches the infrared device should be an environmentally-degraded version of the original thermal energy generated by heat-sources. Because the environmental factors that influence infrared images are complicated, infrared images may show different distributions for the same temperature. We propose that however complex environmental factors are, they stay the same within the same image. Therefore, pixels in the same picture should have the same proportion factor to the original thermal radiation, which we crudely approximated with the black-body radiation (the Stefan-Boltzmann law). The Stefan-Boltzmann law shows that the radiation coming out of a black-body per unit area is proportional to the biquadrate of its absolute temperature:

$$\mathcal{R}(T) = \sigma T^4, \tag{1}$$

where T denotes the absolute temperature of the black body, $R(T)$ denotes its per-unit radiation, and the proportionality coefficient σ is a constant value. Based on Equation 1, the pixel values in an infrared image can be expressed as:

$$\begin{aligned} I(x,y) &= \mathcal{F}(T_{x,y}|w,d,h,v,p,w,UV) \\ &= \mathcal{R}(T_{x,y}) * \mathcal{E}(w,d,h,v,p,w,UV) \\ &= \sigma T_{x,y}^4 * \mathcal{E}(w,d,h,v,p,UV), \end{aligned} \tag{2}$$

where w, d, h, v, p, UV denote the environmental factors that may affect the transmission coefficient, i.e. the weather, dew point, humidity, visibility, pressure and UV index. Note that for the pixel positions (x, y) locating in the same infrared image with pixel value $I(x, y)$, the environmental factors should be fixed, and then there is $I(x, y) \propto T_{x,y}^4$. Consider human beings as constant-temperature entities with the mean body temperature of $T_p = 37.2°C = 310.35K$. Then the mean pixel value I_p for person areas in an infrared image can be used as an anchor point to calculate the temperature map T.

$$T(x,y) = (I(x,y) * T_p^4 / I_p)^{1/4} - 273.15, \tag{3}$$

where T_p is fixed to be the average human body temperature of $310.35K$ and I_p can be computed in each testing image using the primary detection as prior. Figure 3 shows some examples of our computed temperature maps and their corresponding infrared images. With Equation 3, we are able to transform infrared images under various conditions into unified temperature maps. We prove that these temperature maps can help describe pedestrians from a novel aspect, improving the overall detection performance. During training, we calculate the temperature maps with annotations: computing I_p by averaging over all labeled pedestrian areas. During testing, we use a primary detector to extract its most confident pedestrian box in each image for the human pixel-value prior (I_p).

C. Infrared Pedestrian Detection

We show our overall detection framework in Figure 2. In total, there are two detection stages.

Stage I: Primary Detection. To obtain the temperature map on a test image, we first need to know at least one pedestrian position as prior. In this work, we use a primary detector to do this. This primary detection network has the same structure as the temperature net, only that it treats infrared images the same way as gray-scale visible images during training and testing. We test the accuracy of its most-confident detection box on the whole test set of FLIR dataset. The high precision (97.96%) for its most-confident detection proves that the primary detector is capable of providing pedestrian pixel value (I_p).

Stage II: Temperature Map Detection. With the confident detection box ($bbox = [x_0, y_0, w, h]$) which is considered as human area, we are able to obtain the corresponding temperature map of the test image with Equation 3 and the I_p value:

$$I_p = \sum_{x=x_0}^{x+w} \sum_{y=y_0}^{y+h} I(x,y)/(w*h). \quad (4)$$

Based on the computed temperature map, we train a temperature net to re-recognizes pedestrian areas with a more clean form of temperature information, e.g. mean human body temperature fixed to a constant value, relieving influence of environmental variables.

After stage II, we fuse the results from both stages as the final detection. We validate that both stages have its role and that the temperature net largely boosts the overall performance in Section III-C.

D. Implementation Details

During training, we augment images from the training set of the FLIR dataset [40] via flipping. For the primary detector, we regard infrared images as gray-scale visible images, train the network with Stochastic Gradient Descent (SGD) optimizer, each batch containing one image and 128 regions of interest. For better initialization, the network is pre-trained on the COCO dataset for object detection task. We set the initial learning rate to be 1e-3 and decrease by half every 30'000 iterations for a total of 90'000 iterations. For the temperature net, we train on converted temperature maps of the training images, with the same SGD optimizer and the batch size of one image and 128 regions. This time, we initialize from weights of the primary detector, and train with a learning rate of 1e-4 for 30'000 iterations. The whole process is carried out on one TitanX GPU with 12G memory, and it takes 9 hours and 3 hours for training primary and temperature detector, respectively.

III. Experiments

We carry out extensive analysis and experiments on the OTCBVS [41] dataset and the FLIR dataset [40] to validate the proposed algorithm.

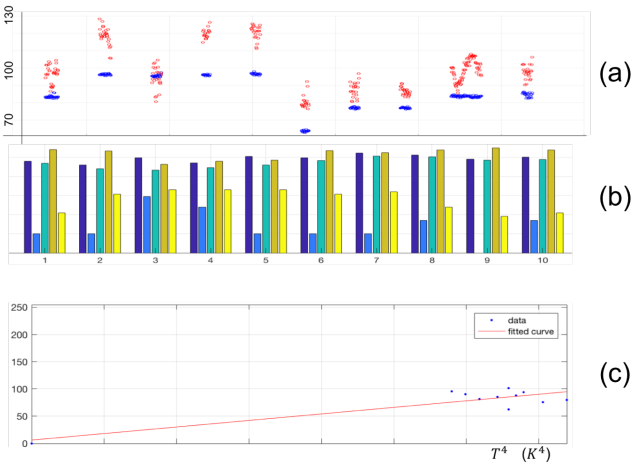

Fig. 4. Data distribution on the OTCBVS dataset. Figure (a) shows the average pixel value for pedestrian (red) and background (blue) areas, respectively. Figure (b) presents the environmental factors with numerical values in the ten capture sessions (bars from left to right represent the environmental factors of air temperature, UV index, dew point, humidity and visibility, respectively). Figure (c) illustrates the fitted curve for averaged pixel value versus the biquadrate of the absolute air temperature.

A. Database and Evaluation Metric

The FLIR dataset [40] is a recently proposed (in June, 2018), large-scale infrared dataset for autonomous driving. Its images are taken from scenes of real-world driving situations (day and night), including annotations for common road objects like pedestrians, cars, bicycles, dogs and etc. The images are of resolution 640 x 512, with a manually-labeled bounding box and category identity of each object, but have no annotations for environmental factors. We choose the class of pedestrian in this dataset to train and test our detection network. In total, we use 5'838 images with 22'372 pedestrians as training set and 1'206 images with 5'779 pedestrians as test set.

The OTCBVS [41] dataset contains 284 challenging infrared images from 10 capture sessions (taken in several days, morning and afternoon, three under rainy weather). The images are with a unified resolution of 360 x 240, taken from a university campus walkway, and have an average of 3-4 persons per image. Each pedestrian is manually labeled with a tight bounding box, and each image has the annotations for a wide range of environmental conditions, i.e. the weather, humidity. We use this dataset to validate our temperature formula (Equation 3), and show that the temperature maps relieve environmental impacts on infrared images. We directly apply our detection network (without fine-tuning) on the OTCBVS dataset to validate the robustness of the proposed method.

We evaluate all our methods with a commonly-used criterion in detection: the precision-recall curve, where detection box that has IoU (intersection over union) larger than 0.5 with ground-truth box is regarded as true positive detection. We also compute the AUC (area-under-curve) value for each precision-

TABLE I

ABLATION STUDY. THIS TABLE SHOWS THE PERFORMANCE OF THE PROPOSED METHOD WITHOUT TEMPERATURE NET (-TNET), WITHOUT ADDING BACK PRIMARY DETECTION BOXES (-PRIDETS), AND THE PRIMARY DETECTION WITHOUT PRE-TRAINING ON THE COCO DATASET (-TNET-COCO). AUC, P AND R DENOTE THE AREA UNDER CURVE, PRECISION AND RECALL VALUES FOR THE OPTIMAL FSCORE POINT (SEE DETAILS IN SECTION III-C), RESPECTIVELY.

	OTCBVS			FLIR		
	AUC	P	R	AUC	P	R
Ours (final)	**0.926**	**0.946**	**0.814**	**0.866**	**0.828**	**0.812**
- TNET	0.888	0.882	0.793	0.823	0.796	0.770
- PRIDETS	0.847	0.907	0.736	0.674	0.786	0.574
- TNET - COCO	0.791	0.990	0.587	0.787	0.729	0.737

TABLE II

COMPARISON WITH METHODS ON THE OTCBVS DATASET. NOTE THAT #HUMAN, #TP AND #FP DENOTE THE NUMBER OF PEDESTRIANS, TRUE POSITIVE DETECTION BOXES AND FALSE POSITIVE DETECTION BOXES, RESPECTIVELY.

No.	#Human	#TP			#FP		
		[42]	[43]	Ours	[42]	[43]	Ours
1	91	88	**90**	77	0	0	3
2	100	94	95	**99**	0	0	2
3	101	**101**	**101**	64	1	1	90
4	109	107	**108**	107	1	0	7
5	101	90	95	**97**	0	0	16
6	97	93	**94**	92	0	0	8
7	94	92	**93**	78	0	0	8
8	99	75	80	**89**	1	1	8
9	95	**95**	**95**	91	0	0	4
10	97	**95**	**95**	91	3	3	18

recall curve for comparisons.

B. Data Analysis

In this section, we analyze the environmental factors in the OTCBVS dataset to better illustrate our motivation of using temperature map in detection.

Environmental factors. As the OTCBVS dataset has images from 10 capture sessions; each session contains labels for some environmental factors, like the weather, humidity, etc. We show the numerical ones with a histogram in Figure 4 (b), where bars from left to right in the same session denote the environmental factors of air temperature, UV index, dew point, humidity and visibility, respectively. To testify that the infrared images are largely affected by these environmental factors, we compute the average pixel value in each image of these 10 sessions. As illustrated in Figure 4 (a), each image is represented by a pair of red and blue points, whose values are the averaged pixel value for pedestrian area and background area, respectively. Note that we managed to align Figure 4 (a) with their session conditions in Figure 4 (b). We can see that images in the same session tend to have similar mean background value, which may be related to their common air temperature. However, images in different sessions appear to have diverse distributions, e.g. the pedestrian and background values are sometimes wide apart and sometimes mixed up, and both of them do not sting to keep to fixed values. This accords with our observation that infrared images are influenced by external environmental factors, and therefore their pixel values do not directly represent temperatures.

Temperature versus pixel value. To validate that infrared pixel values are proportional to the overall radiation which can be calculated with Equation 1, we explore the correlation between the biquadrate of absolute air temperature and the mean pixel values in ten sessions of the OTCBVS dataset. As shown in Figure 4 (c), we draw a linear fitted curve for these two values. We can see that most data points are near the fitted curve, validating the linear relation for infrared image pixel values and the absolute temperature.

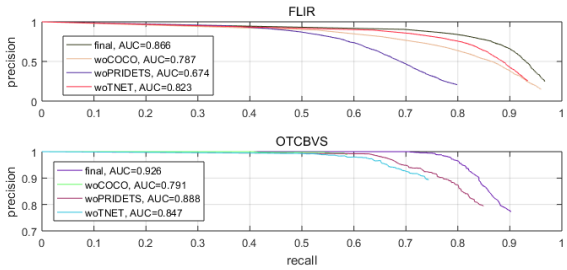

Fig. 5. Precision-recall curves. This figure shows the detailed precision-recall curves for different ablation settings as explained in Table I.

C. Infrared Pedestrian Detection

We carry out an ablation study on the FLIR dataset and the OTCBVS dataset. Note that our networks are trained on the FLIR training set and directly tested on the whole OTCBVS dataset along with the FLIR test set. We compare the final detection performance with ablation settings of:

- TNET: results from the primary detector without incorporating the temperature net;
- PRIDETS: final results without adding back the primary detector outputs;
- TNET & - COCO: results from the primary detector trained with ImageNet classification weights instead of COCO object detection.

Table I shows the evaluation results for these settings in form of AUC (area under the precision-recall curve as shown in Figure 5) and the optimal pair of precision-recall values (P for precision and R for recall) chosen by maximum the F-score ($F = 2PR/(P+R)$). We can see that on the OTCBVS dataset, temperature net alone has better performance than the primary detector; while on the more complicated FLIR dataset, primary detector is better at finding persons. This shows that the primary detector which treats infrared images as gray-scale visible images and the temperature net which is temperature-sensitive specialize in different cognition domain, and that they can complement each other in infrared recognition tasks. The removal of COCO pre-training costs severe loss of AUC on both datasets. This may be due to that the COCO dataset provides a large amount of object detection level annotations and can induce the network to learn strong features with detection priors. On both datasets, the final detection result consistently outperforms any ablation setting, demonstrating

that each part of our proposed framework is significant. Detailed precision-recall curves are drawn in Figure 5. It illustrates the precision-recall trade-offs during detection; it also shows that the proposed method can achieve high recall (about 95%) or high precision (99% or so) for different needs.

We also compare our results with state-of-the-art methods on the OTCBVS dataset [42], [43] in terms of true positive detection number and false positive detection number in each session. Table II shows the overall comparison. We notice a failure in session 3 whose temperature distribution varies a lot from others. Overall, the proposed method performs competitively with methods designed for the OTCBVS dataset, demonstrating its robustness. Although we do not train the network on the OTCBVS dataset, our method still has the best ability at finding pedestrians in some sequences.

IV. Conclusions

In this paper, we propose a novel infrared pedestrian detection framework which places emphasis on the temperature information. To extract temperature information from infrared images, we establish an infrared-temperature transformation formula which can approximately convert the original infrared image into a unified temperature map with the help of a primary pedestrian detection box. We show that the transformed temperature maps relieve environmental impacts and better reveals the heat-source information than the original infrared image. We also show that the temperature information recognizes infrared images from a new aspect, and boosts the overall performance. We demonstrate the effectiveness of our proposed detection module on the FLIR dataset and show its robustness and general applicability on the OTCBVS dataset. In addition, we propose that our infrared-temperature transformation formula can be easily adapted and applied to other infrared-related tasks.

References

[1] M. Bertozzi, A. Broggi, P. G, and etc., "Pedestrian detection in infrared images," in *Intelligent Vehicles Symposium*, 2003. 1

[2] Yajun Fang, K. Yamada, and etc., "A shape-independent method for pedestrian detection with far-infrared images," *IEEE Transactions on Vehicular Technology*, vol. 53, pp. 1679–1697, 2004. 1

[3] H. S. Dong Xia and Z. Shen, "Real-time infrared pedestrian detection based on multi-block lbp," in *International Conference on Computer Application and System Modeling*, 2010. 1

[4] D. Olmeda, A. de la Escalera, and J. M. Armingol, "Detection and tracking of pedestrians in infrared images," in *International Conference on Signals, Circuits and Systems*, 2009. 1

[5] A. Ziebinski, R. Cupek, H. Erdogan, and S. Waechter, "A survey of adas technologies for the future perspective of sensor fusion," in *Computational Collective Intelligence*, 2016. 1

[6] H. Nanda and L. Davis, "Probabilistic template based pedestrian detection in infrared videos," in *Intelligent Vehicle Symposium*, 2002. 1

[7] J. Cheng, Y.-H. Tsai, S. Wang, and M.-H. Yang, "Segflow: Joint learning for video object segmentation and optical flow," in *ICCV*, 2017. 1

[8] M. Yasuno, N. Yasuda, and M. Aoki, "Pedestrian detection and tracking in far infrared images," in *CVPR Workshop*, 2004. 1

[9] Fengliang Xu, Xia Liu, and K. Fujimura, "Pedestrian detection and tracking with night vision," *IEEE Transactions on Intelligent Transportation Systems*, vol. 6, no. 1, pp. 63–71, March 2005. 1

[10] R. OMalley, E. Jones, and M. Glavin, "Detection of pedestrians in far-infrared automotive night vision using region-growing and clothing distortion compensation," *Infrared Physics Technology*, vol. 53, no. 6, pp. 439–449, 2010. 1

[11] Y. Luo, J. Remillard, and D. H, "Pedestrian detection in near-infrared night vision system," in *Intelligent Vehicles Symposium*, 2010. 1

[12] Y. Yang, C. Zhang, J. Lu, and H. Zhang, "Classification of methods in the sins/cns integration navigation system," *IEEE Access*, pp. 3149–3158, 2018. 1

[13] C. Zhang, X. Li, S. Gao, T. Lin, and L. Wang, "Performance analysis of global navigation satellite system signal acquisition aided by different grade inertial navigation system under highly dynamic conditions," *Sensors*, vol. 17, 2017. 1

[14] Y. Gui, P. Guo, H. Zhang, Z. Lei, X. Zhou, J. Du, and Q. Yu, "Airborne vision-based navigation method for uav accuracy landing using infrared lamps," *Journal of Intelligent & Robotic Systems*, Nov 2013. 1

[15] B. W. Li Zhang and R. Nevatia, "Pedestrian detection in infrared images based on local shape features," in *CVPR*, 2007. 1

[16] T. Ojala and I. Harwood, "A comparative study of texture measures with classification based on feature distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996. 1

[17] N. D. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005. 1

[18] V. John, S. Mita, Z. Liu, and B. Qi, "Pedestrian detection in thermal images using adaptive fuzzy c-means clustering and convolutional neural networks," in *International Conference on Machine Vision Applications*, 2015. 1

[19] J. Liu, S. Zhang, S. Wang, and D. N. Metaxas, "Multispectral deep neural networks for pedestrian detection," *CoRR*, 2016. 1

[20] V. A. Knyaz, O. Vygolov, V. V. Kniaz, Y. Vizilter, V. Gorbatsevich, T. Luhmann, and N. Conen, "Deep learning of convolutional auto-encoder for image matching and 3d object reconstruction in the infrared range," in *ICCV Workshops*, 2017. 1

[21] D. Konig, M. Adam, C. Jarvers, G. Layher, H. Neumann, and M. Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *CVPR Workshops*, 2017. 1

[22] C. F. Lin, C. S. Chen, W. J. Hwang, C. Y. Chen, C. H. Hwang, and C. L. Chang, "Novel outline features for pedestrian detection system with thermal images," *Pattern Recognition*, vol. 48, no. 11, pp. 3440–3450, 2015. 1

[23] Hangil Choi, S. Kim, Kihong Park, and K. Sohn, "Multi-spectral pedestrian detection based on accumulated object proposal with fully convolutional networks," in *ICPR*, 2016. 1

[24] A. Khellal, H. Ma, and Q. Fei, "Pedestrian classification and detection in far infrared images," 2015, pp. 511–522. 1

[25] Y.-L. Hou, Y. Song, X. Hao, Y. Shen, M. Qian, and H. Chen, "Multispectral pedestrian detection based on deep convolutional neural networks," *Infrared Physics Technology*, vol. 94, pp. 69 – 77, 2018. 1

[26] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: An astounding baseline for recognition," in *CVPR Workshops*, 2014. 2

[27] R. Girshick, "Fast r-cnn," in *CVPR*, 2015. 2

[28] S. Ren, K. He, R. Girshick, and S. Jian, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis Machine Intelligence*, 2017. 2, 3

[29] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017. 3

[30] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," *CoRR*, 2017. 3

[31] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *CVPR*, 2018. 3

[32] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?" in *CVPR*, 2018. 3

[33] S. Rota Bul, L. Porzi, and P. Kontschieder, "In-place activated batchnorm for memory-optimized training of dnns," in *CVPR*, 2018. 3

[34] Y. Tokozume, Y. Ushiku, and T. Harada, "Between-class learning for image classification," in *CVPR*, 2018. 3

[35] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregation network for instance segmentation," in *CVPR*, 2018. 3

[36] T. Mordan, N. Thome, M. Cord, and G. Hénaff, "Deformable part-based fully convolutional network for object detection," *CoRR*, 2017. 3

[37] S. M. Azimi, E. Vig, R. Bahmanyar, M. Körner, and P. Reinartz, "Towards multi-class object detection in unconstrained remote sensing imagery," *CoRR*, 2018. 3

[38] J. Pang, K. Chen, J. Shi, H. Feng, W. Ouyang, and D. Lin, "Libra r-cnn: Towards balanced learning for object detection," in *CVPR*, 2019. 3

[39] B. F. Jones and P. Plassmann, "Digital infrared thermal imaging of human skin," *IEEE Engineering in Medicine and Biology Magazine*, pp. 41–48, Nov 2002. 3

[40] F. A. Team, "Free flir thermal dataset for algorithm training," https://https://www.flir.asia/oem/adas/dataset/. 4

[41] J. Davis and M. Keck, "Deep residual learning for image recognition," in *Workshop on Applications of Computer Vision*, Jan. 2005. 4

[42] M. Yasuno, N. Yasuda, and M. Aoki, "Pedestrian detection and tracking in far infrared images," in *Intelligent Transportation Systems*, 2004. 5, 6

[43] D. Wu, J. Wang, W. Liu, J. Cao, and Z. Zhou, "An effective method for human detection using far-infrared images," in *First International Conference on Electronics Instrumentation & Information Systems*, 2017. 5, 6