MULTISPECTRAL FUSION FOR OBJECT DETECTION WITH CYCLIC FUSE-AND-REFINE BLOCKS

Heng ZHANG[1,3]*, Elisa FROMONT*[1,4]*, Sébastien LEFEVRE*[2]*, Bruno AVIGNON*[3]

[1]Univ Rennes, IRISA [2]Univ Bretagne Sud, IRISA, [3]ATERMES,[4] IUF, Inria

ABSTRACT

Multispectral images (e.g. visible and infrared) may be particularly useful when detecting objects with the same model in different environments (e.g. day/night outdoor scenes). To effectively use the different spectra, the main technical problem resides in the information fusion process. In this paper, we propose a new halfway feature fusion method for neural networks that leverages the complementary/consistency balance existing in multispectral features by adding to the network architecture, a particular module that cyclically fuses and refines each spectral feature. We evaluate the effectiveness of our fusion method on two challenging multispectral datasets for object detection. Our results show that implementing our *Cyclic Fuse-and-Refine* module in any network improves the performance on both datasets compared to other state-of-the-art multispectral object detection methods.

***Index Terms*—** Multispectral object detection, Multispectral feature fusion, Deep learning

1. INTRODUCTION

Visible and thermal image channels are expected to be complementary when used for object detection in the same outdoor scenes. In particular, visible images tend to provide color and texture details while thermal images are sensitive to objects' temperature, which may be very helpful at night time. However, because they provide a very different view of the same scene, the features extracted from different image spectra may be inconsistent and lead to a difficult, uncertain and error-prone fusion (Fig. 1). In this figure, we use a Convolutional Neural Network (CNN, detailed later) to predict two segmentation masks based on the two (aligned) mono-spectral extracted features from the same image and then fuse the features to detect pedestrians in the dataset. During the training phase, the object detection and the semantic segmentation losses are jointly optimised (the segmentation ground truths are generated according to pedestrian bounding box annotations). We can observe that most pedestrians are visible either on the RGB or on the infrared segmentation masks which illustrates the *complementary* of the channels. However, even though the visible-thermal image pairs are well aligned, the similarity between the two predicted segmentation masks is small, i.e., the multispectral features may be *inconsistent*.

Fig. 1. Examples of thermal and RGB images of the same aligned scenes taken from KAIST multispectral pedestrian detection dataset [1] with detected bounding boxes. The segmentation masks (2nd and 4th columns) are predicted based on the (mono-)spectral features before any fusion process.

In order to augment the consistency between features of different spectra, we design a novel feature fusion approach for convolutional neural networks based on *Cyclic Fuse-and-Refine* modules. Our main idea is to *refine* the mono-spectral features with the *fused* multispectral features multiple times consecutively in the network. Such a fusion scheme has two advantages: 1) since the fused features are generally more discriminative than the spectral ones, the refined spectral features should also be more discriminative than the original spectral features and the fuse-and-refine loop gradually improves the overall feature quality; 2) since the mono-spectral features keep being refined with the same features, their consistency progressively increases, along with the decrease of their complementary, and the consistency/complementary balance is achieved by controlling the number of loops.

We review the related works on multispectral feature fusion with CNN in Section 2. We detail our novel network module named *Cyclic Fuse-and-Refine*, which loops on the fuse-and-refine operations to adjust the multispectral features' complementary/consistency balance in Section 3. In Section 4, we show experiments on the well known KAIST multispectral pedestrian detection dataset [1] on which we obtain new

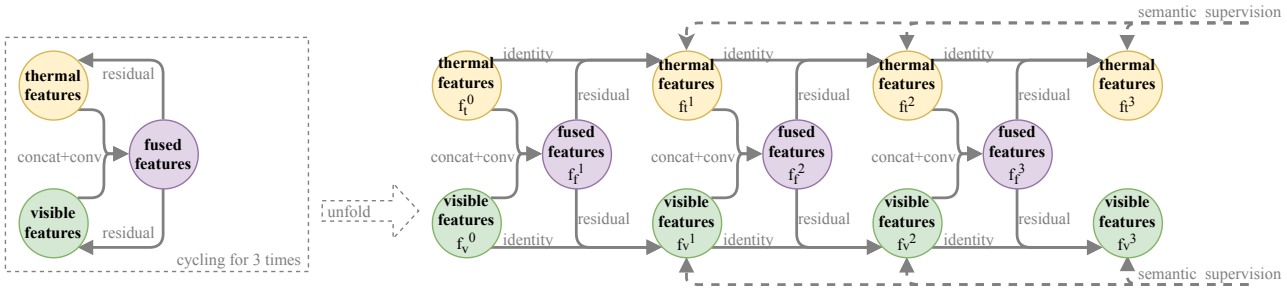

Fig. 2. Illustration (folded on the left part and unfolded on the right) of the proposed *Cyclic Fuse-and-Refine Module* with 3 loops. Better viewed in color.

state-of-the-art results, and on the less known FLIR ADAS dataset [2] on which we set a first strong baseline.

2. RELATED WORK

Existing approaches mainly differ on the strategies ("when" and "how") used to fuse the multispectral features.

When to fuse. The first study on CNN-based multispectral pedestrian detection is made by [3], and they evaluate two fusion strategies: early and late fusions. Then [4] and [5] explore this further and show that a fusion of features halfway in the network, achieves better results than the early or the late fusion. Since then, the halfway fusion has become the default strategy in deep learning-based multispectral (and multimodal) works ([5, 6, 7, 8, 9]). We also choose to locate our fuse-and-refine fusion module halfway in the network.

How to fuse. Features extracted from each spectral channel have different physical properties and choosing how to fuse these complementary information is another central research topic. Basic fusion methods include element-wise addition/average, element-wise maximum and concatenation sometimes in addition to a 1×1 convolution to compress the number of channels as done e.g. in [10]. Building on this, more advanced methods such as [5] and [6] use illumination information to guide the multispectral feature fusion. [11] apply Gated Fusion Units (GFU) [12] to combine two SSD networks [13] on color and thermal inputs. [8] propose a cross-modality interactive attention network to dynamically weight the fusion of thermal/visible features. Our strategy is different: we suggest a cyclic fusion scheme to progressively improve the quality of the spectral features and automatically adjust the complementary/consistence balance.

3. PROPOSED APPROACH

Overview. The fusion and refinement operations are the main ones of our proposed approach. They are repeated (through a cycle) multiple times to increase the consistency of the multispectral features and to decrease the complementarity of the features. An illustration of our *Cyclic Fuse-and-Refine* module with 3 loops in the cycle is presented in Fig. 2.

Fuse-and-Refine. In each loop i, for the fused ($_f$), visible ($_v$) and thermal ($_t$) features, the multispectral feature fusion can be formalized as $f_f^i = \mathcal{F}(\sigma(f_t^{i-1}, f_v^{i-1}))$, where σ is a feature concatenation operation, and $\mathcal{F}$ is a 3×3 convolution followed by a batch normalization operation. For simplicity and to avoid over-fitting, the operation $\mathcal{F}$ in all loops shares weights. The fused features are then assigned as residuals of the spectral features for refinement: $f_t^i = \mathcal{H}(f_t^{i-1} + f_f^i)$, $f_v^i = \mathcal{H}(f_v^{i-1} + f_f^i)$. $\mathcal{H}$ is the activation function (e.g. ReLU).

Semantic supervision. In order to prevent the vanishing gradient problem when learning the parameters of the network and to better guide the multispectral feature fusion, an auxiliary semantic segmentation task is used to bring separate supervision information for each refined spectral features. Concretely, after being refined with the fused features, the thermal and visible features go through a 1×1 convolution (aiming at replacing a fully-connected layer so to ensure a fully-convolutional network) to predict two pedestrian segmentation masks, one for each channel. These predicted masks are also used to tune (or at least visualize) the number of loops in the cyclic module according to the complementary/consistency variations in the features.

Final fusion. Following [14], since the optimal cycling number is unknown and could be different for different image pairs, we aggregate all the refined spectral features to generate the final fused features that will be used for the object detection part of the network. The aggregation is a simple element-wise average function. Let I be the number of loops, the final computation is: $\frac{1}{2I}(\sum_1^I f_t^i + \sum_1^I f_v^i)$.

4. EXPERIMENTS

We evaluate the proposed *Cyclic Fuse-and-Refine Module* on KAIST Multispectral Pedestrian Detection [1] and FLIR ADAS dataset [2], and compare our results with the state-of-the-art multispectral methods. Examples of image pairs with their ground truth bounding boxes are shown in Fig. 3.

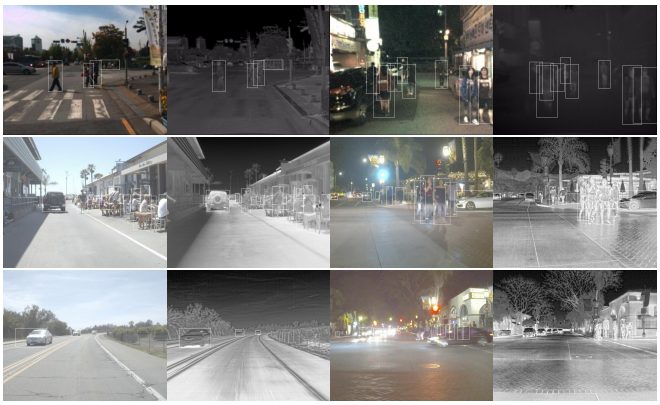

Fig. 3. Examples of visible/thermal image pairs with their ground truth from the KAIST dataset (in the first line) and from the FLIR dataset in the second and third lines (the ground truth annotations are given according to the thermal images). The third line gives an example of misaligned pairs in the FLIR dataset. Better viewed in color and zoomed in.

4.1. Datasets

KAIST. We use the processed version of this multispectral pedestrian detection dataset which contains 7,601 color-thermal image pairs for training and 2,252 pairs for testing. We kept the bounding boxes annotated as "person", "person?" or "people" as positive pedestrian examples. [7] proposed a "sanitized" version of the training annotations which eliminated some of the annotation errors from the original training annotations. According to [4], inaccurate annotations in the test set leads to unfair comparisons, so we only use their "sanitized" testing annotations for our evaluation, with the usual "Miss Rate" performance metric under reasonable setting, i.e., a test subset containing not/partially occluded pedestrians which are larger than 55 pixels.

FLIR. This recently released multispectral (multi-)object detection dataset contains around 10k manually-annotated thermal images with their corresponding reference visible images, collected during daytime and nighttime. We only kept the 3 more frequent classes which are "bicycle", "car" and "person". We manually removed the misaligned visible-thermal image pairs and ended with 4,129 well-aligned image pairs for training and 1,013 image pairs for test [1]. Some examples of the well-aligned and misaligned visible-thermal image pairs are shown in Figure 3.

4.2. Training details

Network architecture. We implemented our *Cyclic Fuse-and-Refine* module on the single stage object detector FSSD [15], which is an improved version of the well known SSD object detector [13]. Note that our proposed module is independent from the chosen network architecture. Following [4] and [5], the mono-spectral features are extracted independently through a VGG16 [16] network, and fused after the conv4_3 layer (halfway through the network). Our baseline architecture uses the element-wise average for the multispectral feature fusion and we integrate and evaluate the proposed module with different number of loops.

Data augmentation. As implemented in SSD [13] and FSSD [15], a few data augmentation methods are applied, such as image random cropping, padding, flipping and distorting for both visible and thermal images.

Anchor designing. Following [17], the anchor designing strategy is adapted for the pedestrian detection for KAIST dataset: we fix the aspect ratio of each anchor box to 0.41 and we only keep three detection layers with scales 32 and $32\sqrt{2}$, 64 and $64\sqrt{2}$, 128 and $128\sqrt{2}$ from fine to coarse respectively. For FLIR, we use the same scale settings but we augment the aspect ratio setting to $\{1, 2, \frac{1}{2}\}$.

Loss functions. To improve object detection, SDS RCNN [18] and MSDS RCNN [7] use an additional task, semantic segmentation, and jointly optimize the loss for the segmentation and detection tasks while training the network. To fairly compare our work to these competitors, we also use this auxiliary loss to supervise the training of the proposed module.

4.3. Comparison with state-of-the-art methods

On KAIST. We compare the experimental results of our approach with state-of-the-art methods in Table 1. For these experiments, we make 3 loops in the Fuse-and-Refine cycle. Depending on what was done in the literature and to allow a fair comparison, we report our detection accuracy with sanitized and original training annotations respectively. All the deep learning-based methods [4, 19, 5, 6, 7] use the same input image resolution (640×512) and the same backbone network (VGG16). The results show that our proposed method allows us to obtain better detection results than all its competitors for both the sanitized and original training annotations. Note that the computational overhead from *CFR* is quite small. During inference, each cycle only add $\sim$0.4ms of inference time.

On FLIR. Because of the misalignment problems in the dataset, there is, to our knowledge, no paper which uses the FLIR dataset [2] for multispectral object detection. We use our sanitized version of the dataset and compare the mAP percentage of two different models: a baseline model which uses the traditional halfway fusion architecture (with the VGG backbone) and the same model with our proposed module. Again, we can see in Table 2 that our method provides important mAP gains for all the considered object categories.

4.4. Ablation study

We study in details (on the KAIST dataset with the sanitized training annotations and the reasonable test set) the ef-

[1] This new aligned dataset can be downloaded here: `http://shorturl.at/ahAY4`

Fig. 4. Examples of pedestrian segmentation masks predicted on 2 visible/thermal image pairs (one taken at day time, one taken at night) of the KAIST dataset after a different number of loops (1-3) in the fuse-and-refine cycle.

fectiveness of the proposed fusion module and the relationship between the number of loops in the fuse-and-refine cycle and the multispectral feature complementary/consistency balance. The experimental results are summarised in Table 3. We provide the Miss Rate and DICE scores [20] between the pedestrian masks predicted by each version of the refined thermal/visible features. These DICE scores are used as an indicator of similarity between the spectral features. From the table we observe successive accuracy gains from the baseline (no loop) to 3 loops, and a decrease after 4 loops; meanwhile the value of DICE scores continue to increase along with the number of loops. We then visualize, on two sample image pairs, the pedestrian masks predicted by visible/thermal features after each refinement in Figure 4. The first column corresponds to input images marked with the detected pedestrians; The second, third and fourth columns correspond to segmentation masks predicted after 1 to 3 loops. The first and third lines (resp. second and fourth) are for visible (res. thermal) images and their corresponding segmentation masks. It can be observed that the quality and similarity of the masks gradually increase with the number of loops. With the increase of similarity between the spectral features, their consistency increases and their complementarity decreases. As mentioned in Section 1, the lack of consistency between the multispectral features is harmful; on the contrary, too much consistency leads to sharp emerge/plunge in the feature values, and makes the fusion meaningless. That explains why the Miss Rate starts to decrease after 4 loops. In practice the number of loops should be tuned for any dataset but we believe that very few values should be tried (between 2 and 5).

Methods	Miss Rate (lower, better)		
	R-All	R-Day	R-Night
Training with sanitized annotations:			
MSDS-RCNN [7]	7.49%	8.09%	5.92%
CFR_3	**6.13%**	**7.68%**	**3.19%**
Training with original annotations:			
ACF+T+THOG [1]	47.24%	42.44%	56.17%
Halfway Fusion [4]	26.15%	24.85%	27.59%
Fusion RPN+BF [19]	16.53%	16.39%	18.16%
IAF R-CNN [5]	16.22%	13.94%	18.28%
IATDNN+IASS [6]	15.78%	15.08%	17.22%
MSDS-RCNN [7]	11.63%	10.60%	13.73%
CFR_3	**10.05%**	**9.72%**	**10.80%**

Table 1. Detection accuracy comparisons in terms of Miss Rate percentage on KAIST Dataset [1]. Our competitors' results are taken from [5] and [7].

Methods	mAP	Bicycle	Car	Person
Baseline	71.17%	56.39%	83.90%	73.28%
CFR_3	**72.39%**	**57.77%**	**84.91%**	**74.49%**

Table 2. mAP results for two CNN object detection architectures which use (or not) our Cyclic Fuse-and-Refine (CFR) blocks on FLIR dataset [2].

Methods	Miss Rate	DICE Scores
Baseline	7.68%	-
CFR_1	6.90%	{64.53%}
CFR_2	6.40%	{78.89%, 89.70%}
CFR_3	6.13%	{74.60%, 90.60%, 94.17%}
CFR_4	7.09%	{58.25%, 85.91%, 92.9%, 96.11%}

Table 3. Miss rates versus DICE scores w.r.t. different numbers of Fuse-and-Refine loops. Each experiment is repeated five times and we report the average performance.

5. CONCLUSION

This paper proposes a novel *cycle fuse-and-refine* module to improve the multispectral feature fusion while taking into account the complementary/consistency balance of the features. Experiments on KAIST [1] and FLIR [2] datasets show that integrating the proposed fusion module to a "vanilla" multispectral pedestrian detector leads to substantial accuracy improvements. Several visible/thermal image pairs have a misalignment problem in FLIR dataset. This problem could be more serious in real world applications due to calibration errors or temporal shifts. A Region Feature Alignment (RFA) module [21] tackled such a cross-modality disparity problem in a supervised manner and in a two-stage object detection setting. In the future, we would like to explore a more general solution to this problem with a similar cyclic-align scheme.

6. REFERENCES

[1] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon, "Multispectral pedestrian detection: Benchmark dataset and baselines," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[2] "Free flir thermal dataset for algorithm training," `https://www.flir.com/oem/adas/adas-dataset-form/`.

[3] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke, "Multispectral pedestrian detection using deep fusion convolutional neural networks," in *24th European Symposium on Artificial Neural Networks, ESANN 2016, Bruges, Belgium, April 27-29, 2016*, 2016.

[4] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas, "Multispectral deep neural networks for pedestrian detection," in *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*, 2016.

[5] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang, "Illumination-aware faster R-CNN for robust multispectral pedestrian detection," *Pattern Recognition*, vol. 85, pp. 161–171, 2019.

[6] Dayan Guan, Yanpeng Cao, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang, "Fusion of multispectral data through illumination-aware deep neural networks for pedestrian detection," *Information Fusion*, vol. 50, pp. 148–157, 2019.

[7] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang, "Multispectral pedestrian detection via simultaneous detection and segmentation," in *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, 2018, p. 225.

[8] Lu Zhang, Zhiyong Liu, Shifeng Zhang, Xu Yang, Hong Qiao, Kaizhu Huang, and Amir Hussain, "Cross-modality interactive attention network for multispectral pedestrian detection," *Information Fusion*, vol. 50, pp. 20–29, 2019.

[9] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu, "Weakly aligned cross-modal learning for multispectral pedestrian detection," in *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[10] Min Lin, Qiang Chen, and Shuicheng Yan, "Network in network," arXiv:1312.4400, 2013.

[11] Yang Zheng, Izzat H. Izzat, and Shahrzad Ziaee, "GFD-SSD: gated fusion double SSD for multispectral pedestrian detection," *CoRR*, vol. abs/1903.06999, 2019.

[12] John Edison Arevalo Ovalle, Thamar Solorio, Manuel Montes y Gómez, and Fabio A. González, "Gated multimodal units for information fusion," *CoRR*, vol. abs/1702.01992, 2017.

[13] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "Ssd: Single shot multibox detector," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016.

[14] Zhiwen Fan, Huafeng Wu, Xueyang Fu, Yue Huang, and Xinghao Ding, "Residual-guide network for single image deraining," in *Proceedings of the 26th ACM International Conference on Multimedia*, New York, NY, USA, 2018, MM '18, p. 1751–1759, Association for Computing Machinery.

[15] Zuoxin Li and Fuqiang Zhou, "FSSD: feature fusion single shot multibox detector," *CoRR*, vol. abs/1712.00960, 2017.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[17] Liliang Zhang, Liang Lin, Xiaodan Liang, and Kaiming He, "Is faster r-cnn doing well for pedestrian detection?," arXiv:1607.07032, 2016.

[18] Garrick Brazil, Xi Yin, and Xiaoming Liu, "Illuminating pedestrians via simultaneous detection & segmentation," in *Proceedings of the IEEE International Conference on Computer Vision*, Venice, Italy, 2017.

[19] Daniel König, Michael Adam, Christian Jarvers, Georg Layher, Heiko Neumann, and Michael Teutsch, "Fully convolutional region proposal networks for multispectral person detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017, pp. 243–250.

[20] Lee R. Dice, "Measures of the amount of ecologic association between species," *Ecology*, vol. 26, no. 3, pp. 297–302, 1945.

[21] Lu Zhang, Zhiyong Liu, Xiangyu Chen, and Xu Yang, "The cross-modality disparity problem in multispectral pedestrian detection," *CoRR*, vol. abs/1901.02645, 2019.