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

Heng ZHANG¹,³, Elisa FROMONT¹,⁴, Sébastien LEFEVRE², Bruno AVIGNON³

¹Univ Rennes, IRISA ²Univ Bretagne Sud, IRISA, ³ATERMES, ⁴IUF, 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 as shown in Figure 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.

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 multisp-
nal pedestrian detection dataset [1] on which we obtain new 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

Multispectral object detection 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]–[7], [8], [9]). We also choose to locate our fuse-and-refine 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/consistency 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.

Fusion. In each loop i, the fused (f), visible (v) and thermal (t) features, the multispectral feature fusion can be formalized as $f_i = F(\sigma(f_i^{t-1}, f_i^{v-1}))$, where $\sigma$ is a feature concatenation operation, and $F$ is a 3 × 3 convolution followed by a batch normalization operation. For simplicity and to avoid over-fitting, the Conv-BN operations $F$ in all loops share the same weights. The fused features are then assigned as residuals of the spectral features for refinement: $f_i^t = H(f_i^{t-1} + f_i^v)$, $f_i^v = H(f_i^{v-1} + f_i^t)$. $H$ 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 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. 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: $\frac{1}{2I}(\sum_{i=1}^{I} f_i^t + \sum_{i=1}^{I} f_i^v)$.

4. EXPERIMENTS

We evaluate the proposed Cyclic Fuse-and-Refine Module on KAIST Multispectral Pedestrian Detection [11] and FLIR ADAS dataset [2], and compare our results with the state-of-the-art multispectral methods. Examples of image pairs from these datasets with their ground truth bounding boxes are shown in Figure 2.
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. [2] 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 [5]. 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 [14], 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 [15] 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 [14], 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 [16], 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 $\{1, 2, 4\}$.

**Loss functions.** To improve object detection, SDS RCNN [17] 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 [4,18,5,6,7] 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.

**On FLIR.** Because of the misalignment problems in the dataset, there is, to our knowledge, no paper which uses the FLIR dataset 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 effectiveness of the proposed fusion module and the relation-
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. Zoom in to see details.

ship 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 [19] 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 (resp. 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 emergence/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).

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

<table>
<thead>
<tr>
<th>Methods</th>
<th>Miss Rate (lower, better)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>R-All</td>
</tr>
<tr>
<td>MSDS-RCNN [7]</td>
<td>7.49%</td>
</tr>
<tr>
<td>CFRM_3</td>
<td>6.13%</td>
</tr>
</tbody>
</table>

### Table 2. mAP results for two CNN object detection architectures which use (or not) our cyclic fuse-and-refine module (CFRM) on FLIR dataset [2].

<table>
<thead>
<tr>
<th>Methods</th>
<th>mAP</th>
<th>Bicycle</th>
<th>Car</th>
<th>Person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>71.17%</td>
<td>56.39%</td>
<td>83.90%</td>
<td>73.28%</td>
</tr>
<tr>
<td>CFRM_3</td>
<td>72.39%</td>
<td>57.77%</td>
<td>84.91%</td>
<td>74.49%</td>
</tr>
</tbody>
</table>

### 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.

<table>
<thead>
<tr>
<th>Methods</th>
<th>Miss Rate</th>
<th>DICE Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>7.68%</td>
<td></td>
</tr>
<tr>
<td>CFRM_1</td>
<td>6.90%</td>
<td>{64.53%}</td>
</tr>
<tr>
<td>CFRM_2</td>
<td>6.40%</td>
<td>{78.89%, 89.70%}</td>
</tr>
<tr>
<td>CFRM_3</td>
<td>6.13%</td>
<td>{74.60%, 90.60%, 94.17%}</td>
</tr>
<tr>
<td>CFRM_4</td>
<td>7.09%</td>
<td>{58.25%, 85.91%, 92.9%, 96.11%}</td>
</tr>
</tbody>
</table>

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 was proposed by [20] to tackle such 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


