Weakly Misalignment-free Adaptive Feature Alignment for UAVs-based Multimodal Object Detection

Chen Chen Jiahao Qi Xingyue Liu Kangcheng Bin Ruigang Fu Xikun Hu Ping Zhong*
National University of Defense Technology, China
chenchen21c@nudt.edu.cn, qijiahao1996@nudt.edu.cn, liuxingyue18@nudt.edu.cn,
binkc21@nudt.edu.cn, furuigang08@nudt.edu.cn, xikun@nudt.edu.cn, zhongping@nudt.edu.cn

Abstract

Visible-infrared (RGB-IR) image fusion has shown great potentials in object detection based on unmanned aerial vehicles (UAVs). However, the weakly misalignment problem between multimodal image pairs limits its performance in object detection. Most existing methods often ignore the modality gap and emphasize a strict alignment, resulting in an upper bound of alignment quality and an increase of implementation costs. To address these challenges, we propose a novel method named Offset-guided Adaptive Feature Alignment (OAFA), which could adaptively adjust the relative positions between multimodal features. Considering the impact of modality gap on the cross-modality spatial matching, a Cross-modality Spatial Offset Modeling (CSOM) module is designed to establish a common subspace to estimate the precise feature-level offsets. Then, an Offset-guided Deformable Alignment and Fusion (ODAF) module is utilized to implicitly capture optimal fusion positions for detection task rather than conducting a strict alignment. Comprehensive experiments demonstrate that our method not only achieves state-of-the-art performance in the UAVs-based object detection task but also shows strong robustness to the weakly misalignment problem.

1. Introduction

Visible-infrared (RGB-IR) object detection on unmanned aerial vehicles (UAVs) has attracted extensive attention [2, 27, 39], due to its ability to make full use of complementary information and achieve robust around-the-clock detection. However, the collected RGB-IR images are typically mismatched in resolutions and field of views owning to disparate imaging principles. Thus, it is challenging to fuse the corresponding information of target between multimodal images.

*Corresponding author.

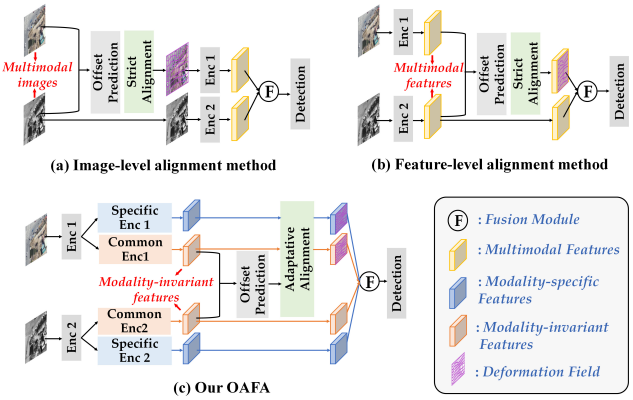


Figure 1. Previous multimodal object detection architectures under weakly misalignment conditions vs. our OAFA. (a) Image-level alignment method: executing a pixel-level strict alignment by explicitly extracting the global deformation fields from multimodal images. (b) Feature-level alignment method: explicitly estimating feature offsets through multimodal features to achieve strict feature alignment. It can be further classified into global and local methods based on whether alignment is exclusively performed on target regions. (c) Our OAFA: the basic global offsets are estimated by the modality-invariant features to eliminate the impact of modality gap. Based on this offset, we perform adaptive feature alignment without strict alignment.

To increase the quality of fusion representations, researchers always apply pre-registration techniques, involving image cropping [43] and affine transformation [47] before the fusing process. Nevertheless, since distinct imaging time in multimodal sensors and intricate movements in targets, it is difficult, or even impossible, to achieve precise UAVs image alignment [29]. It inevitably leads to local inconsistencies in multimodal spatial distribution, named the *weakly misalignment* problem [41], resulting in a mismatch of the multimodal target features in corresponding positions [40]. Since the detection task is sensitive to spatial locations, the weakly misalignment problem would lead to significant performance degradation. Several studies have been carried out to deal with this issue by accomplishing

a strict alignment based on offset estimation, as shown in Fig. 1a and Fig. 1b. Although these methods have made significant progress, there still remain two main challenges:

Inaccurate offset estimation. Conventional methods devote to predicting spatial offsets by identifying the distribution and appearance discrepancies in image or feature level. However, in fact, the discrepancies arise not only from spatial offsets but also from modality gap [31], which stems from the differences in imaging principles and wavelengths between RGB and IR sensors. This gap could make it hard to correctly match cross-modality representations in spatial, leading to misleading offset estimation. Hence, the first challenge is *how to bridge the modality gap to achieve accurate cross-modal offset estimation*.

High dependency on strict alignment. The strict alignment is necessary for most existing methods. Unfortunately, the strict alignment always requires additional annotation information, such as the ground truth offsets between multimodal representations [45]. It will lead to higher implementation costs that limit real-world applications. Moreover, the detection performance relies heavily on the quality of alignment. However, even the state-of-the-art alignment methods have the risk of failure in some scenarios [18], which may cause the error delivery to the detection task. Consequently, the second challenge is *how to enhance the robustness of the detection model to the weakly misalignment problem without strict alignment*.

To this end, the purpose of this paper is to achieve adaptive alignment between RGB (sensed modality) and IR (reference modality) images based on spatial offsets for superior detection performance instead of strict alignment. Technically, we propose a novel method named Offset-guided Adaptive Feature Alignment (OAFA) for UAVs-based multimodal object detection under weakly misalignment conditions. It mainly consists of a Cross-modality Spatial Offset Modeling (CSOM) module and an Offset-guided Deformable Alignment and Fusion (ODAF) module for dealing with the two challenges, respectively. Considering the first challenge, CSOM establishes a cross-modal common subspace to obtain the modality-invariant features for eliminating modality gap. As distributional aligned mappings, the modality-invariant features could be utilized to evaluate spatial consistency between multimodal features for more accurate cross-modal offsets estimation.

As for the second challenge, ODAF is designed to adaptively adjust feature positions by implicit offset compensation, aiming to make full use of the offsets and achieve superior detection performance without strict alignment. In specific, ODAF involves a learnable deformation component that dynamically predicts the offsets of sampling convolution kernels. With the help of the offsets provided by CSOM, ODAF captures optimal sampling positions in sensed modality features to attain comprehensive target representations after fusion. This module predicts the corresponding sampling positions for diverse regions in different image pairs, contributing to a strong flexibility of OAFA in handling diverse and complex offsets. Besides, the training stability of deformable components is improved since the explicit guidance of the offsets reduces the burden of the adaptive alignment.

Our contributions can be summarized as follows: (1) We propose a robust RGB-IR object detection method for UAVs images under weakly misalignment conditions, effectively implementing adaptive feature alignment instead of strict alignment. (2) Considering the difficulty of estimating cross-modality offset, we build a multimodal common subspace to reduce the impact of modality gap on multimodal spatial matching. (3) An offset-guided deformable alignment module is exploited to achieve adaptive alignment by implicitly capturing optimal fusion positions, enhancing the accuracy of the detection task. (4) Extensive experiments on public dataset show that our OAFA is superior to state-of-the-art models only with a simple fusion strategy. Moreover, it proves the effectiveness in addressing the weakly misalignment problem.

2. Related Work

Multimodal object detection under weakly misalignment conditions. A general solution to address the weakly misalignment problem is to conduct pixel-level pre-alignment [24, 46, 52], which learn a global mapping function from multimodal images to correct the pixel coordinates. Many of these methods [3, 16, 26] are supervised with the deformable fields in order to achieve strict pre-alignment. To achieve end-to-end training, Zhang *et al*. [45] first introduced a data-driven region feature alignment module to predict proposal regions offsets from multimodal features, which was indispensably constrained by the target locations in RGB-IR images. On this basis, Yuan *et al*. [41] made full use of the oriented multimodal annotations to realize multimodal target alignment in position, size, and angle. Due to the requirement of extra supervisions for the strict alignment process, the above methods are labor-intensive and costly to implement. Besides, the existing modality gap at both the pixel-level and feature-level leads to incorrect estimation of spatial offsets. Consequently, it is indeed significant to propose a novel method for RGB-IR object detection under weakly misalignment conditions.

Common subspace learning. Common subspace learning aims to exploit shared information among multiple data. In field of multimodal learning, it can be used to bridge the modality gap by extracting the unified latent representations in the inter-modality samples. Works can be broadly grouped into transformation-based [9, 23, 32, 37], weight-sharing-based [12, 14, 15], and decoupling-based methods [13, 20, 38]. **The transformation-based methods**

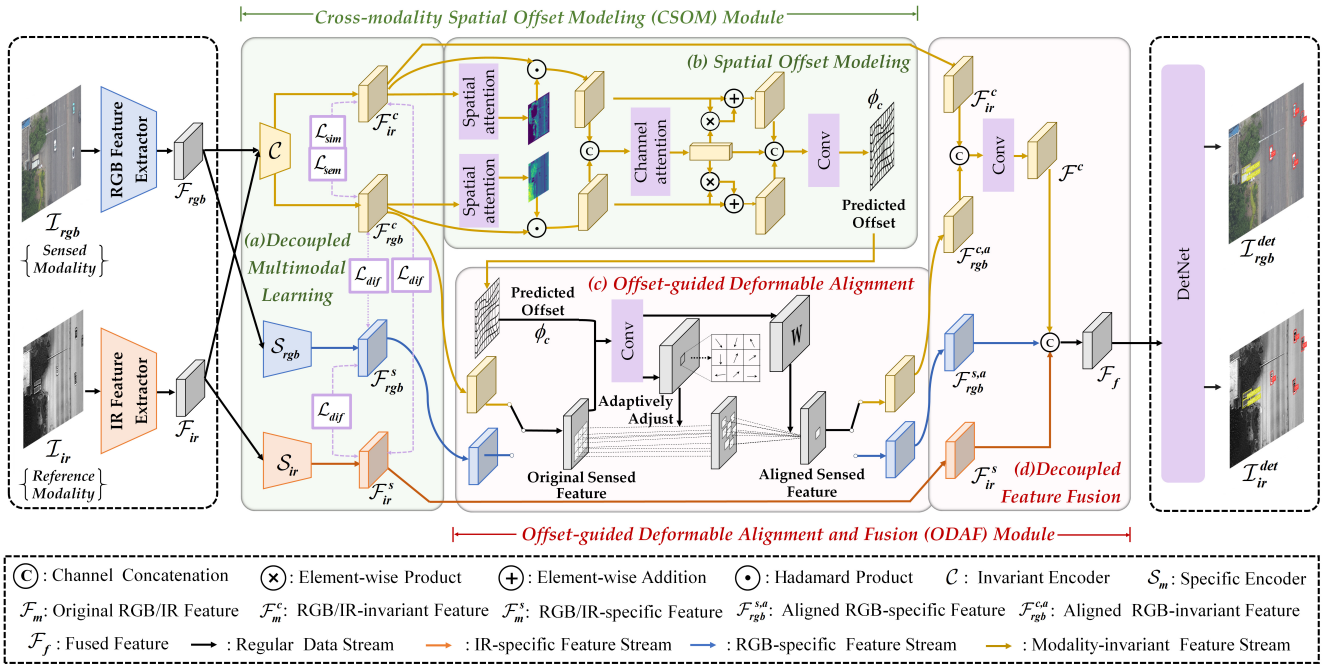


Figure 2. An Overview of OAFA. It consists of two components: a Cross-modality Spatial Offset Modeling (CSOM) module and an Offset-guided Deformable Alignment and Fusion (ODAF) module. In our implementation, we choose the RGB images as the sensed modality whose features are shifted to achieve alignment, and the IR images as the reference modality. The CSOM can decouple the original multimodal features to modality-invariant and modality-specific features and predict a superior spatial offset in a common subspace. The ODAF utilizes this basic offset to adaptively align the sample positions of the RGB features with IR features, and then fuses the aligned RGB and original IR features. Finally, the fused features are conducted to generate the ultimate detection results.

generally transform one modality (images/features) into another modality by domain transfer [9] or auto-encoder [23, 32, 37]. It is hard to achieve an end-to-end training due to the requirement of a pre-trained transformation network. **The weight-sharing-based methods** devote to designing the weight-shared network within a two-stream framework to only capture modality-invariant features [12, 14, 15]. Nevertheless, overemphasis on common features may result in the underutilization of complementary information. **The decoupling-based methods** overcome this obstacle by trying to separate multimodal features into modality-invariant and modality-specific features with a series of constraint functions [13, 20, 38]. Inspired by the third category methods, we develop a decoupled multimodal learning network with three constraints for the detection task. Moreover, due to the complexity of UAVs-based multimodal features, we design the network with partial weight sharing and custom encoder architecture to achieve effective disentanglement.

Deformable convolution. Convolutional neural networks (CNNs) are inherently limited in modeling geometric transformations because of their fixed kernel structures. To enhance the transformation modeling capability of CNNs, Dai *et al*. [8] designed the deformable convolution, which has shown superior performance in several high-level vision tasks [1, 28, 51]. In recent years, it has also been used in the field of video super-resolution [5, 30, 33, 34]. As a pioneer work, Tian *et al*. [30] adopted the deformable convolution to align the features of adjacent frames in the temporal domain. It demonstrates that deformable convolution can effectively capture the motion cues of the targets. Nonetheless, Chan *et al*. [4] has proved that the deformable alignment is difficult to train. Training instability frequently leads to offset overflow, degrading the final performance in downstream task. Therefore, we construct a basic offset as explicit guidance to reduce the burden of deformable alignment training.

3. Method

3.1. Overview

Our method aims to enhance the accuracy and robustness of the RGB-IR object detection model under weakly misalignment conditions. As shown in Fig. 2, a pair of weakly misaligned RGB-IR images are fed into a two-stream network to acquire the multiscale multimodal representations. Subsequently, a decoupled learning network is designed to develop a common subspace to figure out the modality-invariant and modality-specific features. To mitigate the influence of modality gap on offset estimation, we exclusively employ the modality-invariant features to predict spatial offsets. Then, this basic offsets are utilized to implic-

itly adjust the spatial positions of the modality-invariant and modality-specific features. It accomplishes adaptive feature alignment that is tailored for RGB-IR object detection task without any extra supervisions. Finally, the aligned RGB and original IR features are fused with a concatenation operation to generate final detection results.

3.2. Cross-modality Spatial Offset Modeling Module

Given the RGB features F_{rgb} and IR features F_{ir}, CSOM predicts the spatial offsets ϕ_c as basic knowledge for the subsequent alignment module. The CSOM consists of a decoupled multimodal learning submodule and a spatial offset modeling submodule.

Decoupled multimodal learning. This submodule mainly builds a common subspace $\tilde{\mathbb{C}} = [\mathbb{C}_{rgb}, \mathbb{C}_{ir}]$ and a specific subspace $\tilde{\mathbb{S}} = [\mathbb{S}_{rgb}, \mathbb{S}_{ir}]$, as shown in Fig. 2a. For each input RGB-IR image pair $\mathcal{I}_{rgb} \in \mathbb{R}^{H\times W\times 3}$ and $\mathcal{I}_{ir} \in \mathbb{R}^{H\times W\times 1}$, we map it to the multimodal representations $\mathcal{F}_m (m \in \{rgb, ir\})$ by a basic pyramid feature extractor $\mathcal{B}_m$. The $\mathcal{F}_m$ is the bounded random samples on the interval $[a, b]^N$. Then, the modality-invariant features $\mathcal{F}_m^c$ and modality-specific features $\mathcal{F}_m^s$ projected from $\tilde{\mathbb{C}}$ and $\tilde{\mathbb{S}}$ are figured out by the encoding functions:

$$\mathcal{F}_m^s = \mathcal{S}_m(\mathcal{F}_m; \theta_m^s),\ \mathcal{F}_m^c = \mathcal{C}_m(\mathcal{F}_m; \theta^c), \tag{1}$$

where the invariant feature encoder $\mathcal{C}_m$ shares the parameters θ^c across two modalities, whereas the specific feature encoder $\mathcal{S}_m$ assigns separate parameters θ_m^s for each modality. Considering the diversity information in UAVs images, we not only design the decoupled network in objective function but also in network architecture.

In terms of the network architecture, an accepted assumption is that the input features are more related in the low layers, while the high-layer features tend to be uncorrelated [48]. Thus, $\mathcal{C}_m$ is constructed with a convolutional layer followed by the SiLU activation function. In contrast, $\mathcal{S}_m$ contains a convolutional layer and a C3 module with three convolutional and a series of bottleneck layers to obtain deeper features.

As for the objective functions, inspired by [13], we apply three types of constraints for decoupled multimodal learning: *Similarity Loss, Difference Loss, and Semantic Loss*. First, we employ a similarity loss $\mathcal{L}_{sim}$ to minimize the distance between $\mathcal{F}_{rgb}^c$ and $\mathcal{F}_{ir}^c$ to learn common subspace $\tilde{\mathbb{C}}$. The $\mathcal{L}_{sim}$ is defined by a state-of-the-art distance metric named central moment discrepancy [42] to measure the distribution dissimilarity between two features:

$$\begin{aligned}\mathcal{L}_{sim} =& \frac{1}{|b-a|}\left\|E(\mathcal{F}_{rgb}^c) - E(\mathcal{F}_{ir}^c)\right\|_2 \\ &+ \sum_{k=2}^{K}\frac{1}{|b-a|^k}\left\|C_K(\mathcal{F}_{rgb}^c) - C_K(\mathcal{F}_{ir}^c)\right\|_2,\end{aligned} \tag{2}$$

where $E(\cdot)$ is the empirical expectation vector, $C_K(\cdot)$ is the vector of k-th order sample central moments, and $\|\cdot\|_2$ represents the Euclidean norm.

Second, we present a difference loss $\mathcal{L}_{dif}$ for highlighting the dissimilarity between $\mathcal{F}_{rgb}^s$ and $\mathcal{F}_{ir}^s$ to capture complementary information. Additionally, as $\mathcal{F}_m^c$ and $\mathcal{F}_m^s$ are expected to be mutually independent, $\mathcal{L}_{dif}$ should be extended to them as well. These discrepancies are achieved by the squared Frobenius norm:

$$\begin{aligned}\mathcal{L}_{dif} =& \sum_{m\in\{rgb,ir\}}\sqrt{Tr(((F_m^s)^T F_m^c)^T((F_m^s)^T F_m^c))} \\ &+ \sqrt{Tr(((F_{rgb}^s)^T F_{ir}^s)^T((F_{rgb}^s)^T F_{ir}^s))},\end{aligned} \tag{3}$$

where $Tr(\cdot)$ is trace of the matrix.

Third, only restrained by the above objectives, there exists a potential risk that $\mathcal{C}_m$ learns useless noisy cues, whereas $\mathcal{S}_m$ acquires almost all modality information. To address this issue, we add a semantic loss $\mathcal{L}_{sem}$ to enforce $\mathcal{C}_m$ to learn discriminative features:

$$\mathcal{L}_{sem} = \sum_{m\in\{rgb,ir\}}(\mathcal{L}_{cls}(\mathcal{F}_m^c) + \mathcal{L}_{reg}(\mathcal{F}_m^c) + \mathcal{L}_{obj}(\mathcal{F}_m^c)), \tag{4}$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_{obj}$ are the detection loss functions of the YOLOv5 [17]. It should be noted that this object detection branch, serving as an auxiliary task, is not incorporated during the inference stage to enhance the efficiency of our OAFA. Finally, the total loss of the decoupled multimodal learning submodule can be defined as:

$$\mathcal{L}_{dec} = \lambda_1\mathcal{L}_{sim} + \lambda_2\mathcal{L}_{sim} + \lambda_3\mathcal{L}_{sem}, \tag{5}$$

where λ_i is the trade-off weights. In this study, we set λ_1, λ_2, and λ_3 to 0.03, 0.01, and 0.03 respectively.

Spatial offset modeling. We model the spatial offsets ϕ_c by the estimating spatial differences between $\mathcal{F}_{rgb}^c$ and $\mathcal{F}_{ir}^c$. In the field of change detection [49], the differences are calculated directly by subtracting or concatenating two features. However, it would introduce significant noises when dealing with intricate backgrounds in UAVs images. To solve this problem, a feature difference enhancement module is presented to filter out irrelevant changes and concentrate on shift objects. As shown in Fig. 2b, $\mathcal{F}_m^c$ is firstly fed into a spatial attention module to capture the representations of the target and crucial background areas in each modality. The structure of the spatial attention module is the same as [35]. In this case, the spatially enhanced unimodal feature $F_m^{c,p}$ can be computed as:

$$\mathcal{F}_m^{c,p} = \sigma(f_{7\times7}(Cat(Max(\mathcal{F}_m^c), Mean(\mathcal{F}_m^c)))) \odot \mathcal{F}_m^c, \tag{6}$$

where σ is the sigmoid function, $f_{i\times j}$ denotes the $i \times j$ convolution layer, $Cat(\cdot)$ refers to the concatenation operation, and $\odot$ indicates the Hadamard product operation.

We conduct a concatenation instead of subtraction strategy to extract the changing features F_{dif} to mitigate the influence of complex background. Then, a channel difference enhanced module is introduced to further suppress channel noise. Specifically, we employ simultaneous average-pooling and max-pooling to generate distinct spatial context descriptors. Following this, channel attention can be derived by transmitting the spatial context descriptors to a shared multilayer perceptron network f_{mlp}. $\mathcal{F}_m^{c,p}$ are combined with the channel attention map to get the spatial-channel enhanced features $\mathcal{F}_m^{c,e}$. The above operation is represented as:

$$\begin{aligned}\mathcal{F}_m^{c,e} = \sigma(f_{mlp}(AvgPool(\mathcal{F}_{dif}))+ \\ f_{mlp}(MaxPool(\mathcal{F}_{dif}))) \times \mathcal{F}_m^{c,p} + \mathcal{F}_m^{c,p}.\end{aligned} \tag{7}$$

Finally, ϕ_c is obtained through a fusion of the $\mathcal{F}_{rgb}^{c,e}$ and $\mathcal{F}_{ir}^{c,e}$ along with a nonlinear layer.

3.3. Offset-guided Deformable Alignment and Fusion Module

We have obtained the offsets ϕ_c from CSOM, which represents the spatial differences between $\mathcal{F}_{rgb}$ and $\mathcal{F}_{ir}$. Inspired by [5], the ϕ_c can be an effectively initialized offset to unsupervised alignment task. Hence, ODAF is designed to adaptively capture optimal fusion location with the help of the basic knowledge ϕ_c, incorporating desired complementary information for the subsequent detection task. It is composed of two submodules: offset-guided deformable alignment and decoupled feature fusion.

Offset-guided deformable alignment. We mainly apply a deformable convolution to achieve implicit offset compensation and adaptive alignment for the detection task. Unlike the default deformable convolution that learns offsets of convolution kernels from their original features, the offsets generated from our module are mapped from the basic offsets ϕ_c, which allows for more effective feature correction and reliable model training. Considering different contributions from each offset region for the detection task, we add a modulation scalar Δm_k that is learned from ϕ_c to dynamically aggregate information around the corresponding position p. As shown in Fig. 2c, given the center sampling value $x(p)$ in the sensed features $\mathcal{F}_{rgb}^{n}(n \in \{s, c\})$ and the basic offsets ϕ_c, each counterpart $y(p)$ in aligned features $\mathcal{F}_{rgb}^{n,a}$ can be obtained as follows:

$$y(p) = \sum_{k=1}^{K} \omega_k \cdot x(p + \phi_c + p_k + \Delta p_k) \cdot \Delta m_k, \tag{8}$$

where K, p_k, and w_k denote the number of kernel weights, the k-th fixed offset, and the k-th kernel weight, respectively. For instance, $p_k \in \{(-1, 1), (-1, 0), ..., (1, 1)\}$ donates a regular grid of a 3×3 kernel defined with $K = 9$. Δp_k denotes the implicit compensation to ϕ_c of the k-th location. It is estimated with $conv(x)[p + \phi_c]$ where ϕ_c can be regarded as a prior of Δp_k in the standard deformable convolution. Besides, since the location $p + \phi_c + p_k + \Delta p_k$ may be fractional, we adopt bilinear interpolation to obtain the final sampling values.

Decoupled feature fusion. In traditional fusion process, $\mathcal{F}_m^c$ and $\mathcal{F}_m^s$ are directly concatenated to obtain the fused features $\mathcal{F}_f$ [6]. Unfortunately, the similarity constraint between $\mathcal{F}_{rgb}^{c,a}$ and $\mathcal{F}_{ir}^{c}$ leads to much redundancy. To eliminate duplicate information and increase discriminative representations, $\mathcal{F}_{rgb}^{c,a}$ and $\mathcal{F}_{ir}^{c}$ should be combined and optimized before fusion. Thus, the fused features F_f are defined as:

$$\mathcal{F}_f = Cat(f_{1\times1}(Cat(\mathcal{F}_{rgb}^{c,a}, \mathcal{F}_{ir}^{c})), \mathcal{F}_{rgb}^{s,a}, \mathcal{F}_{ir}^{s}). \tag{9}$$

3.4. Model Training

Training loss. The overall method is based on a supervised detection framework. The purpose of our multimodal alignment is not to form a strict alignment but to achieve optimal adjustment between multimodal features to improve the performance of the subsequent detection task. Thus, the alignment subtask is integrated into the main detection task instead of adding extra alignment loss. The total loss function of our OAFA can be represented as follows:

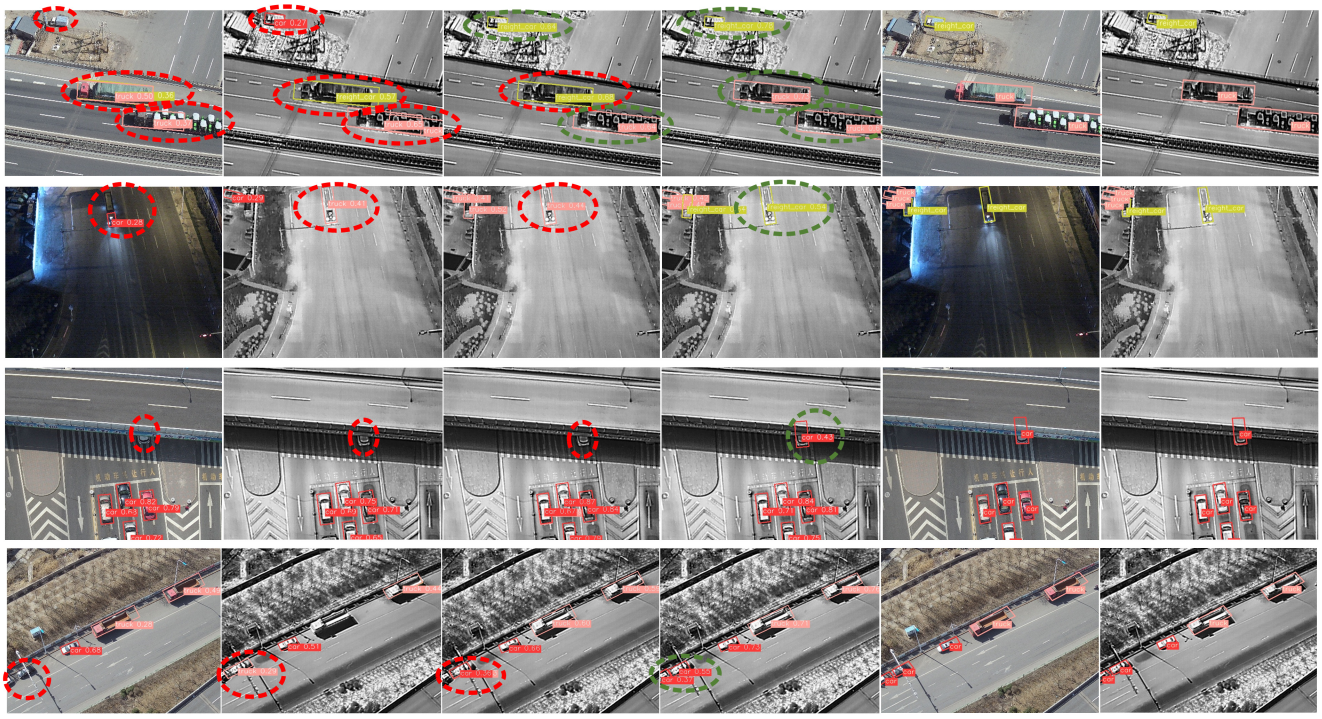
$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{obj} + \mathcal{L}_{dec}. \tag{10}$$

Two-stage training strategy. A premise for effective feature alignment is that $\mathcal{F}_m^c$ contains sufficient spatial information, thereby estimating more accurate basic offsets. To this end, we adopt a two-stage learning strategy to train our OAFA end-to-end. In stage I, to gain effective $\mathcal{F}_m^c$, we try to ensure that pre-decoupling features $\mathcal{F}_m$ contain more comprehensive target representations. Technically, we only train the feature extraction and fusion module with $\mathcal{L}_{cls}$, $\mathcal{L}_{reg}$, and $\mathcal{L}_{obj}$. Specifically, a pair of $\mathcal{I}_{rgb}$ and $\mathcal{I}_{ir}$ are fed into a two-stream network to extract multiscale features $\mathcal{F}_m$ and gain fusion features $\mathcal{F}_f$ with the concatenation operation. Then, the $\mathcal{F}_f$ is put into the detection network to obtain the detection results. In stage II, we apply all the proposed modules to our network. In this process, the unimodal feature extraction network $\mathcal{B}_m$ is initialized with the weights from stage I. Then, the joint optimization of the unimodal feature extraction and multimodal feature alignment network is carried out effectively with $\mathcal{L}_{total}$.

4. Experiments

4.1. Experimental Settings

Datasets and metrics. We evaluate our method on DroneVehicle dataset [29], which is a large-scale multimodal dataset for vehicle detection in UAVs images. It contains 28,439 image pairs with 953,087 vehicles, covering



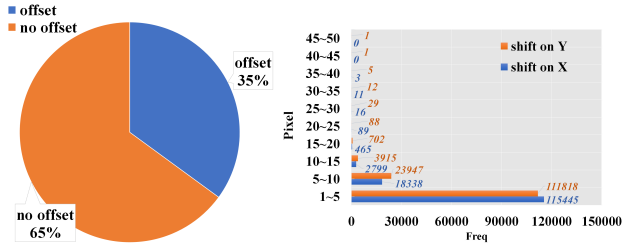
(a) Result in Unimal RGB (b) Result in Unimal IR (c) Result in Baseline (d) Result in Ours (e) Groundtruth in RGB (f) Groundtruth in IR

Figure 3. Four Examples of detection results on the test set of DroneVehicle dataset under weakly misalignment conditions. The confidence threshold is set to 0.25. The results of fusion methods are visual in IR images to correspond to the supervisory label. Red, green, and pink rectangles represent car, freight-car, and truck targets, respectively. Objects correctly detected are represented in green dashed circles, while incorrectly detected objects are represented in red dashed circles.

different viewing angles, heights, lighting conditions, and scenarios. The dense-oriented annotations are made for five categories: car, truck, bus, van, and freight-car. Though the DroneVehicle performs pre-registration on each RGB-IR image pair, the weakly misalignment problem still exists. As shown in Fig. 4, about 35% bounding boxes exhibit deviation issues in DroneVehicle dataset, and the shift distance mostly ranges from 0 to 15 pixels.

This dataset is divided into 17,990 image pairs for training, 1,469 image pairs for validating, and the remaining 8,980 image pairs for testing. Following the standard evaluation metrics, we report the mean average precision (mAP) with intersection over union (IoU) threshold of 0.4 on the validation set.

Training details. The experiments are carried out on a single NVIDIA RTX A6000 GPU with 48 GB of memory. We implement our algorithm with Pytorch toolbox and SGD optimizer with a momentum of 0.937 and a weight decay of 0.0005. The initial learning rate is set to 0.01 and eventually reduced to 0.002 by cosine annealing [22]. The input size of our model is set to 640×640 in the preprocessing stage and the batch size is 16. The training epoch is set to 150 with 50 and 100 epochs in the first and second stages, respectively. We use the ground truth of the IR images as the training label, owning to the more comprehensive target



(a) Statistics of misalignment ratio (b) Statistics of shift value

Figure 4. The statistics of ground truth bounding-boxes misalignment ratio and shift value within RGB-IR image pairs on DroneVehicle dataset.

annotations available in the IR modality.

4.2. Results Comparisons

Compared methods. We compare our method with the state-of-the-art unimodal and multimodal object detection methods. **Unimodal methods:** We choose a series of oriented object detection methods for comparison, including the one-stage methods: RetinaNet [19], S^2A-Net [11], and YOLOv5s [17]; and the two-stage methods: Faster R-CNN [25], Oriented R-CNN [36], and RoITransformer [10]. RGB and IR images are separately used as training data to evaluate the detection performance of each algorithm. **Multimodal fusion methods:** We also compare our method with eight recent fusion methods:

Table 1. Detection results (mAP, in %) on DroneVehicle dataset. Note that all detectors locate and classify vehicles with OBB heads. The speed refers to the speed of network inference without post-processing (batch size = 1). In speed, our method is only compared with the fusion methods. Best results are highlighted in **bold**. And the second one is marked with underline.

Detectors	Modality	Car	Truck	Freight-car	Bus	Van	mAP (%) ↑	Speed (fps) ↑
RetinaNet [19]	RGB	78.5	34.4	24.1	69.8	28.8	47.1	-
Faster R-CNN [25]		79.0	49.0	37.2	77.0	37.0	55.9	-
Oriented R-CNN [36]		80.1	53.8	41.6	85.4	43.3	60.8	-
S^2A-Net [11]		80.0	54.2	42.2	84.9	43.8	61.0	-
RoITransformer [10]		61.6	55.1	42.3	85.5	44.8	61.6	-
YOLOv5s [17]		78.6	55.3	43.8	87.1	46.0	62.1	-
RetinaNet [19]	IR	88.8	35.4	39.5	76.5	32.1	54.5	-
Faster R-CNN [25]		89.4	53.5	48.3	87.0	42.6	64.2	-
Oriented R-CNN [36]		89.8	57.4	53.1	89.3	45.4	67.0	-
S^2A-Net [11]		89.9	54.5	55.8	88.9	48.4	67.5	-
RoITransformer [10]		90.1	60.4	58.9	89.7	52.2	70.3	-
YOLOv5s [17]		90.0	59.5	60.8	89.5	53.8	70.7	-
UA-CMDet [29]	RGB+IR	87.5	60.7	46.8	87.1	38.0	64.0	9.1
Halfway Fusion [21]		90.1	62.3	58.5	89.1	49.8	70.0	20.4
CIAN [44]		90.1	63.8	60.7	89.1	50.3	70.8	21.7
AR-CNN [45]		90.1	64.8	62.1	89.4	51.5	71.6	18.2
MBNet [50]		90.1	64.4	62.4	88.8	53.6	71.9	21.7
TSFADet [41]		89.9	67.9	63.7	89.8	54.0	73.1	18.6
C^2Former [40]		90.2	68.3	64.4	89.8	58.5	74.2	-
SLBAF-Net [7]		90.2	72.0	68.6	89.9	59.9	76.1	**63.2**
Ours		**90.3**	**76.8**	**73.3**	**90.3**	**66.0**	**79.4**	33.1

UA-CMDet [29], Halfway Fusion [21], CIAN [44], MBNet [50], AR-CNN [45], TSFADet [41], C^2Former [40], and SLBAF-Net [7]. Among them, UA-CMDet, MBNet, TSFADet, and C^2Former all consider the weakly misalignment problem during the fusion process, which affirms the universality and significance of this problem. For each compared method, we use their original experimental settings to ensure equity.

Quantitative comparison. The results are shown in Tab. 1. In the unimodal methods, benefitting from a well-designed multiscale feature extraction network, YOLOv5s has comparable detection accuracy (62.1% in RGB images and 70.7% in IR images), even better than some fusion methods. Correspondingly, the multimodal methods SLBAF-Net, constructed upon the two-stream YOLOv5 framework, has also yielded excellent results. Compared with all methods, we can find that our OAFA achieves the best average performance on the validation set, which is 3.3% higher than the second place. It is worth noting that our method has obvious advantages in distinguishing confusing categories (truck and freight-car). The reason may stem from the fact that OAFA excels in identifying the optimal spatial correspondences of multimodal features, obtaining reliable features of the target after fusion.

Qualitative comparison. Our baseline model is SLBAF-Net, a two-stream framework based on YOLOv5s where RGB and IR branches are fused with a concatenation operation in multiscale features. Among the compared methods mentioned above, we select the optimal unimodal YOLOv5s and the baseline model as the qualitative comparative model. We provide visual detection results of the compared methods in Fig. 3. It can be observed that, given the constraint of the IR labels, the objects are successfully detected by OAFA when the target positions are shifted, while both unimodal and baseline model experience classification and localization errors or fail to detect, especially for occluded and few-shot object detection. The underlying causes can be speculated that the weak misalignment problem confuses the fusing model to learn discriminative features. In contrast, our method has the capability to capture aligned and reliable features, improving the performance in classification and localization processes.

Speed comparison. We compare the speed of OAFA with other fusion detection methods to prove the efficiency of our method. For a fair comparison, all the detection models are tested with batch size 1, and the input size is set to 640×640. As shown in Tab. 1, our proposed OAFA achieves a speed of 33.1 FPS, outperforming most existing multimodal object detection methods. Although our method requires more computational time compared to SLBAF-Net, it would be worthwhile to invest this additional time to achieve superior performance. Furthermore, it can be seen that OAFA can accomplish real-time detection on the NVIDIA RTX A6000 platform.

4.3. Ablation Study

In this section, we conduct several ablation studies to validate the effectiveness of the different components in our

proposed method. The results are shown in Tab. 2.

Effect of CSOM module. We first analyze the effect of the CSOM module as it provides the fundamental offset estimation, which determines the stability of the model training. We perform ablation experiments on both the decoupled multimodal learning (DML) module and the spatial offset modeling (SOM) module separately. As presented in the first to third rows of Tab. 2, the performance of the baseline model is improved by 0.4% and 1.6% after incorporating the SOM and DML modules, respectively. It verifies that both coarse-grained feature alignment and increasing fusion flexibility through decouple learning are effective for the detection task. Besides, when integrating these two submodules, the performance improvement of model over baseline is 2.1%, which indicates that the quality of the feature alignment can be enhanced by mitigating modal gap.

Effect of ODAF module. We then verify the necessity of the ODAF module in addressing the weak misalignment problem. The results demonstrate that despite ODAF has poor stability in model training without dependable offset guidance, incorporating ODAF into the baseline can provide +1.7% significant improvement. It is owing to adaptively adjust the optimal sampling position for acquiring reliable information in RGB images.

Effect of two-stage training. The two-stage training strategy provides comprehensive unimodal representations for multimodal alignment and fusion. We can observe that the accuracy decreases to 77.3% if we directly train the whole network without two-stage training. It is 0.9% and 0.5% lower than exclusively incorporating CSOM and ODAF, respectively. The reason could be that the inappropriate initialization in feature extraction network leads to training instability and makes it hard to obtain abundant representations in spatial and semantic. After undergoing a two-stage training process, our method attains a performance level of 79.4%, surpassing the capabilities of any individual module. It demonstrates that the two-stage training strategy effectively enhances training accuracy and stability. The superior method to fully exploit the strengths of CSOM and ODAF will be further studied in future work.

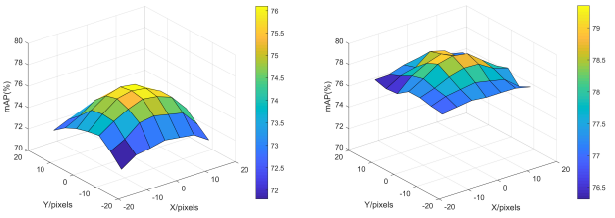
All in all, not only each independent module but also the entire proposed framework can contribute to the multimodal object detection under weakly misalignment conditions.

4.4. Robustness to Position Shift

To quantitatively assess the robustness of our method to the weakly misalignment issue, we conduct experiments on DroneVehicle by manually simulating positional offsets. The tested model is trained on the original training image pairs and validated on the shifted image pairs. In our experiments, we hold the IR images constant while introducing spatial offsets along the x-axis and y-axis for RGB images. The variations in pixel values are defined within the range $\{(\Delta x, \Delta y) \mid \Delta x, \Delta y \in [15, -15]; \Delta x, \Delta y \in \mathbb{Z}\}$. The quantitative outcomes of this experiment in baseline and our model are discernible from Fig. 5. It can be seen that in response to larger position shifts, our method consistently exhibits minor fluctuations in performance, whereas the baseline model experiences a noticeable performance decline. This result demonstrates the robustness of OAFA under weakly misalignment conditions.

Table 2. Ablation study on DroneVehicle dataset. The baseline model is SLBAF-Net, a two-stream framework based on YOLOv5s where RGB and IR branch are fused with a multiscale concatenation operation. The D denotes DML, S denotes SOM, C denotes CSOM, and O denotes ODAF.

Method	CSOM		ODAF	Two-stage Training	mAP (%) ↑
	DML	SOM			
Baseline					76.1
Baseline+S		✓			76.5
Baseline+D	✓				77.7
Baseline+C	✓	✓			78.2
Baseline+O			✓		77.8
Baseline+C+O	✓	✓	✓		77.3
OAFA	✓	✓	✓	✓	**79.4**



(a) SLBAF-Net (Baseline) (b) Our method

Figure 5. Surface plot visualization of the position shift experiments. Horizontal coordinates indicate various shift steps along the x-axis and y-axis. Vertical coordinate denotes the detection performances in terms of mAP.

5. Conclusion

In this paper, we propose a robust RGB-IR object detection method OAFA for UAVs images under weakly misalignment conditions. Different from previous methods, OAFA can achieve adaptive multimodal feature alignment without strict alignment. To mitigate the impact of the modality gap on multimodal spatial matching, the modality-invariant features are acquired in a common subspace to estimate accurate offsets. Taking the offsets as basic knowledge, the adaptive alignment implicitly captures optimal fusion positions and steadily carries out feature-level corrections with a deformable alignment module. Experimental results on DroneVehicle dataset demonstrate the superior performances of OAFA compared with the state-of-the-art methods. Moreover, our method shows robustness against the weakly misalignment problem in position shift experiments.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (62201586).

References

[1] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *ECCV*, pages 342–357, 2018. 3

[2] Ilker Bozcan and Erdal Kayacan. AU-AIR: A multi-modal unmanned aerial vehicle dataset for low altitude traffic surveillance. In *ICRA*, pages 8504–8510, 2020. 1

[3] Xiaohuan Cao, Jianhua Yang, Jun Zhang, Dong Nie, Minjeong Kim, Qian Wang, and Dinggang Shen. Deformable image registration based on similarity-steered CNN regression. In *MICCAI*, pages 300–308, 2017. 2

[4] Kelvin C. K. Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. Understanding deformable alignment in video super-resolution. In *AAAI*, pages 973–981, 2021. 3

[5] Kelvin C. K. Chan, Shangchen Zhou, Xiangyu Xu, and Chen Change Loy. Basicvsr++: Improving video super-resolution with enhanced propagation and alignment. In *CVPR*, pages 5962–5971, 2022. 3, 5

[6] Hao Chen, Yongjian Deng, Youfu Li, Tzu-Yi Hung, and Guosheng Lin. RGBD salient object detection via disentangled cross-modal fusion. *IEEE TIP*, 29:8407–8416, 2020. 5

[7] Xiaolong Cheng, Keke Geng, Ziwei Wang, Jinhu Wang, Yuxiao Sun, and Pengbo Ding. Slbaf-net: Super-lightweight bimodal adaptive fusion network for uav detection in low recognition environment. *Multimedia Tools and Applications*, pages 1–20, 2023. 7

[8] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 3

[9] Chaitanya Devaguptapu, Ninad Akolekar, Manuj M. Sharma, and Vineeth N. Balasubramanian. Borrow from anywhere: Pseudo multi-modal object detection in thermal imagery. In *CVPR*, pages 1029–1038, 2019. 2, 3

[10] Jian Ding, Nan Xue, Yang Long, Gui-Song Xia, and Qikai Lu. Learning roi transformer for oriented object detection in aerial images. In *CVPR*, pages 2849–2858, 2019. 6, 7

[11] Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE TGRS*, 60: 1–11, 2022. 6, 7

[12] Renlong Hang, Zhu Li, Pedram Ghamisi, Danfeng Hong, Guiyu Xia, and Qingshan Liu. Classification of hyperspectral and lidar data using coupled cnns. *IEEE TGRS*, 58(7): 4939–4950, 2020. 2, 3

[13] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. MISA: modality-invariant and -specific representations for multimodal sentiment analysis. In *ACM MM*, pages 1122–1131, 2020. 2, 3, 4

[14] Danfeng Hong, Naoto Yokoya, Jocelyn Chanussot, and Xiao Xiang Zhu. Cospace: Common subspace learning from hyperspectral-multispectral correspondences. *IEEE TGRS*, 57(7):4349–4359, 2019. 2, 3

[15] Danfeng Hong, Jocelyn Chanussot, Naoto Yokoya, Jian Kang, and Xiao Xiang Zhu. Learning-shared cross-modality representation using multispectral-lidar and hyperspectral data. *IEEE GRSL*, 17(8):1470–1474, 2020. 2, 3

[16] Lloyd Haydn Hughes, Diego Marcos, Sylvain Lobry, Devis Tuia, and Michael Schmitt. A deep learning framework for matching of sar and optical imagery. *ISPRS*, 169:166–179, 2020. 2

[17] Glenn Jocher. ultralytics/yolov5. https://github.com/ultralytics/yolov5, Oct. 2020. 4, 6, 7

[18] Huafeng Li, Junzhi Zhao, Jinxing Li, Zhengtao Yu, and Guangming Lu. Feature dynamic alignment and refinement for infrared-visible image fusion: Translation robust fusion. *Information Fusion*, 95:26–41, 2023. 2

[19] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2999–3007, 2017. 6, 7

[20] Fan Liu, Zhiyong Cheng, Huilin Chen, Anan Liu, Liqiang Nie, and Mohan S. Kankanhalli. Disentangled multimodal representation learning for recommendation. *arXiv:2203.05406*, 2022. 2, 3

[21] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. In *BMVC*, 2016. 7

[22] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *ICLR*, 2017. 6

[23] Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *AAAI*, pages 164–172, 2020. 2, 3

[24] Sayan Nag. Image registration techniques: A survey. *arXiv:1712.07540*, 2017. 2

[25] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE TPAMI*, 39(6):1137–1149, 2017. 6, 7

[26] Marc-Michel Rohé, Manasi Datar, Tobias Heimann, Maxime Sermesant, and Xavier Pennec. Svf-net: Learning deformable image registration using shape matching. In *MICCAI*, pages 266–274, 2017. 2

[27] Hazim Shakhatreh, Ahmad Sawalmeh, Ala I. Al-Fuqaha, Zuochao Dou, Eyad Almaita, Issa Khalil, Noor Shamsiah Othman, Abdallah Khreishah, and Mohsen Guizani. Unmanned aerial vehicles (uavs): A survey on civil applications and key research challenges. *IEEE Access*, 7:48572–48634, 2019. 1

[28] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, pages 536–553, 2018. 3

[29] Yiming Sun, Bing Cao, Pengfei Zhu, and Qinghua Hu. Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning. *IEEE TCSVT*, 32(10):6700–6713, 2022. 1, 5, 7

[30] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. TDAN: temporally-deformable alignment network for video super-resolution. In *CVPR*, pages 3357–3366, 2020. 3

[31] Di Wang, Jinyuan Liu, Xin Fan, and Risheng Liu. Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration. In *IJCAI*, pages 3508–3515, 2022. 2

[32] Guan'an Wang, Tianzhu Zhang, Jian Cheng, Si Liu, Yang Yang, and Zengguang Hou. Rgb-infrared cross-modality person re-identification via joint pixel and feature alignment. In *ICCV*, pages 3622–3631, 2019. 2, 3

[33] Hua Wang, Dewei Su, Chuangchuang Liu, Longcun Jin, Xianfang Sun, and Xinyi Peng. Deformable non-local network for video super-resolution. *IEEE Access*, 7:177734–177744, 2019. 3

[34] Xintao Wang, Kelvin C. K. Chan, Ke Yu, Chao Dong, and Chen Change Loy. EDVR: video restoration with enhanced deformable convolutional networks. In *CVPR*, pages 1954–1963, 2019. 3

[35] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: convolutional block attention module. In *ECCV*, pages 3–19, 2018. 4

[36] Xingxing Xie, Gong Cheng, Jiabao Wang, Xiwen Yao, and Junwei Han. Oriented R-CNN for object detection. In *ICCV*, pages 3500–3509, 2021. 6, 7

[37] Han Xu and Jiayi Ma. Emfusion: An unsupervised enhanced medical image fusion network. *Information Fusion*, 76:177–186, 2021. 2, 3

[38] Xing Xu, Kaiyi Lin, Lianli Gao, Huimin Lu, Heng Tao Shen, and Xuelong Li. Learning cross-modal common representations by private-shared subspaces separation. *IEEE Trans. Cybern.*, 52(5):3261–3275, 2022. 2, 3

[39] Huang Yao, Rongjun Qin, and Xiaoyu Chen. Unmanned aerial vehicle for remote sensing applications - A review. *Remote Sensing*, 11(12):1443, 2019. 1

[40] Maoxun Yuan and Xingxing Wei. $\mathbf{C}^2$former: Calibrated and complementary transformer for rgb-infrared object detection. *arXiv:2306.16175*, 2023. 1, 7

[41] Maoxun Yuan, Yinyan Wang, and Xingxing Wei. Translation, scale and rotation: Cross-modal alignment meets rgb-infrared vehicle detection. In *ECCV*, pages 509–525, 2022. 1, 2, 7

[42] Werner Zellinger, Thomas Grubinger, Edwin Lughofer, Thomas Natschläger, and Susanne Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. In *ICLR*, 2017. 4

[43] Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Grid anchor based image cropping: A new benchmark and an efficient model. *IEEE TPAMI*, 44(3):1304–1319, 2022. 1

[44] 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*, 50:20–29, 2019. 7

[45] Lu Zhang, Zhiyong Liu, Xiangyu Zhu, Zhan Song, Xu Yang, Zhen Lei, and Hong Qiao. Weakly aligned feature fusion for multimodal object detection. *arXiv:2204.09848*, 2022. 2, 7

[46] Xinyue Zhang, Chengcai Leng, Yameng Hong, Zhao Pei, Irene Cheng, and Anup Basu. Multimodal remote sensing image registration methods and advancements: A survey. *Remote Sensing*, 13(24):5128, 2021. 2

[47] Zhimeng Zhang and Yu Ding. Adaptive affine transformation: A simple and effective operation for spatial misaligned image generation. In *ACM MM*, pages 1167–1176, 2022. 1

[48] Zixiang Zhao, Haowen Bai, Jiangshe Zhang, Yulun Zhang, Shuang Xu, Zudi Lin, Radu Timofte, and Luc Van Gool. Cddfuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion. In *CVPR*, pages 5906–5916, 2023. 4

[49] Jiaxiang Zheng, Yichen Tian, Chao Yuan, Kai Yin, Feifei Zhang, Fangmiao Chen, and Qiang Chen. Mdesnet: Multitask difference-enhanced siamese network for building change detection in high-resolution remote sensing images. *Remote Sensing*, 14(15):3775, 2022. 4

[50] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *ECCV*, pages 787–803, 2020. 7

[51] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3

[52] Barbara Zitová and Jan Flusser. Image registration methods: a survey. *Image Vis. Comput.*, 21(11):977–1000, 2003. 2