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

Supplementary Material

A. Implementation Details

A.1. Framework of Our OAFA

First, we show the framework of our OAFA:

Algorithm 1 Framework of the Offset-guided Adaptive Feature Alignment

```
Input:
    The image of RGB modality (sensed modality), $\mathcal{I}_{rgb}$;
    The image of IR modality (reference modality), $\mathcal{I}_{ir}$;
    The threshold of output confidence $\tau$;
Output:
    A set of detection results $\mathcal{D}$ on the RGB-IR images;
1: Extracting unimodal feature maps $\mathcal{F}_{rgb}$ and $\mathcal{F}_{ir}$ from $\mathcal{I}_{rgb}$ and $\mathcal{I}_{ir}$;
2: Conducting modality-invariant and modality-specific features $\mathcal{F}^c_m$, $\mathcal{F}^s_m$ ($m \in \{rgb, ir\}$) from $\mathcal{F}_{rgb}$ and $F_{ir}$ through a decoupled multimodal learning network;
3: Predicting the spatial offsets $\phi_c$ from $\mathcal{F}^c_{rgb}$ and $\mathcal{F}^c_{ir}$ using the spatial offset modeling submodule;
4: Capturing optimal fusion locations in sensed modality to gain aligned features $\mathcal{F}^{c,a}_{rgb}$ and $\mathcal{F}^{s,a}_{rgb}$ with the help of the $\phi_c$ by the offset-guided deformable alignment submodule;
5: Fusing the aligned RGB features $\mathcal{F}^{c,a}_{rgb}$, $\mathcal{F}^{s,a}_{rgb}$ and original IR features $\mathcal{F}^c_{ir}$, $\mathcal{F}^s_{ir}$ by the decoupled feature fusion submodule to obtain the final features $\mathcal{F}_f$;
6: Conducting the anchor-wise classification and regression to obtain the detection category $\mathcal{C}a_i$, confidence $\mathcal{C}o_i$, and 2D oriented bounding box coordinates $\mathcal{O}_i$ for each target $\mathcal{T}_i$;
7: $\mathcal{D} \leftarrow \varnothing$;
8: for each $\mathcal{T}_i$ do
9:     if $\mathcal{C}o_i > \tau$ then
10:        $\mathcal{D} \leftarrow \mathcal{D} \cup \{[\mathcal{O}_i, \mathcal{C}a_i, \mathcal{C}o_i]\}$;
11:    end if
12: end for
13: return $\mathcal{D}$;
```

A.2. The Structure of the Spatial Offset Modeling Submodule

The structure of the Spatial Offset Modeling (SOM) submodule is shown in Fig. 1, mainly including a spatial attention network and a channel attention network. The input of SOM is the modality-invariant features $\mathcal{F}^c_{rgb}$ and $\mathcal{F}^c_{ir}$, and the output is the spatial offsets ϕ_c. In spatial attention network, we first use a max and an average operation in channel to generate two channel context descriptors, and reduce the channel dimensions of the input feature maps to 1. Following this, we gain richer representations in textures and backgrounds by concatenating these two feature maps. And then use a convolutional layer and the Sigmoid function to optimize and normalize the feature maps to [0, 1]. Finally, these feature maps are multiplied by the original feature maps to obtain the spatial enhanced feature maps $\mathcal{F}^{c,p}_{rgb}$ and $\mathcal{F}^{c,p}_{ir}$. As for channel attention network, $\mathcal{F}^{c,p}_{rgb}$ and $\mathcal{F}^{c,p}_{ir}$ are added for calculating the spatial differences. After that, the multimodal channel attention weights are computed by a two-layer perceptron following the global max-average pooling operation. The ultimate attention weights are derived by integrating the dual channel attention weights and applying a Sigmoid activation function. Acting as a scale factor for each channel, the transformed attention weights are multiplied element-wise with the original feature maps. The channel-enhanced feature maps $\mathcal{F}^{c,e}_{rgb}$ and $\mathcal{F}^{c,e}_{ir}$ are obtained by adding the original feature maps with the scaled feature maps. Eventually, The spatial offsets ϕ_c are derived from the differences between the $\mathcal{F}^{c,e}_{rgb}$ and $\mathcal{F}^{c,e}_{ir}$.

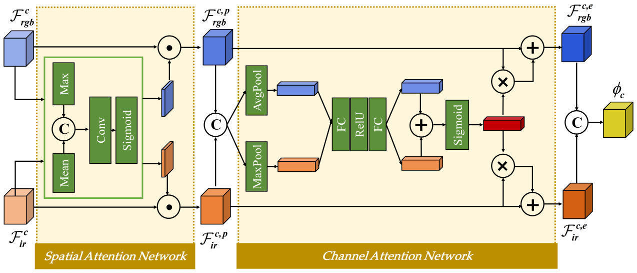


Figure 1. Structure visualization of the spatial offset modeling module.

A.3. The Pipeline of the Two-stage Training

The pipeline of two-stage training is shown in Fig. 2. The purpose of stage I is to ensure that the pre-decoupling features contain sufficient semantic and spatial information. To this end, we train the two-stream multiscale encoder network $\mathcal{B}_m$ to acquire unimodal image features $\mathcal{F}_{rgb}$ and $\mathcal{F}_{ir}$. After that, these features are fed into a concatenation operation to obtain the fusion features, which are constrained by the object detection loss. Note that the pre-trained weights of YOLOv5s [6] are not loaded to avoid the environment bias introduced by its pre-training on natural images. In the training process of stage I, we use an initial learning rate

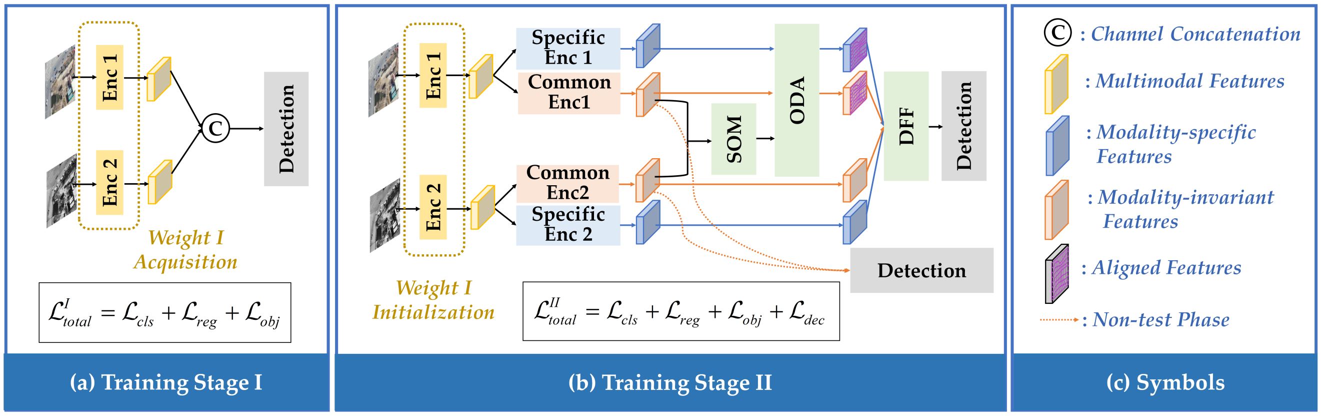
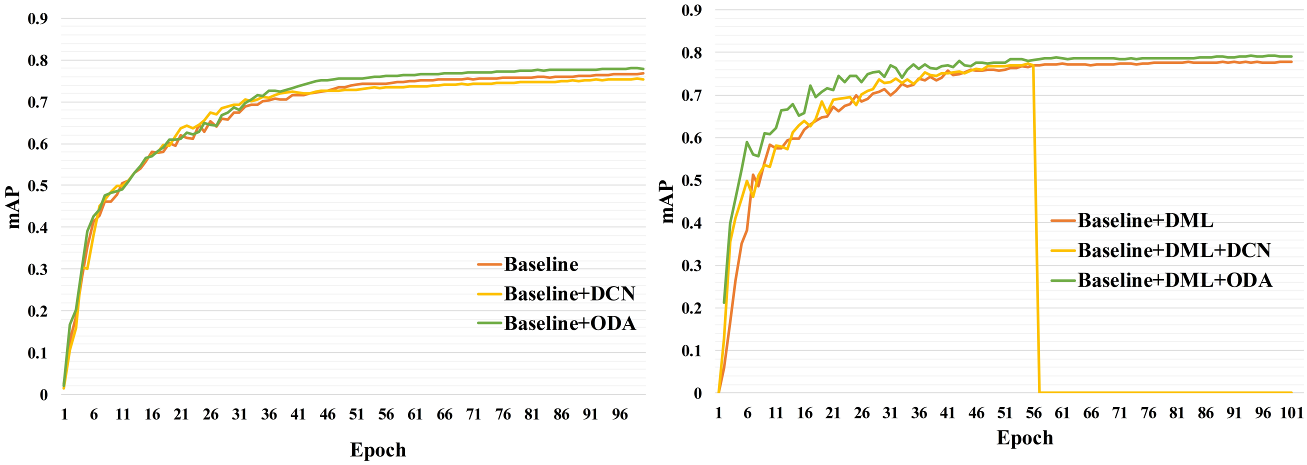

Enc 1
Enc 2
C
Detection
Weight I Acquisition
$\mathcal{L}^{I}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{obj}$
(a) Training Stage I
Enc 1
Enc 2
Specific Enc 1
Common Enc1
Common Enc2
Specific Enc 2
SOM
ODA
DFF
Detection
Detection
Weight I Initialization
$\mathcal{L}^{II}_{total} = \mathcal{L}_{cls} + \mathcal{L}_{reg} + \mathcal{L}_{obj} + \mathcal{L}_{dec}$
(b) Training Stage II
C : Channel Concatenation
: Multimodal Features
: Modality-specific Features
: Modality-invariant Features
: Aligned Features
: Non-test Phase
(c) Symbols

Figure 2. The pipeline of the two-stage training. SOM denotes the spatial offset modeling submodule, ODA denotes the offset-guided deformable alignment submodule, and DFF denotes the decoupled feature fusion submodule.



(a) Results in model without DML

(b) Results in model with DML

Figure 3. Influence of DCN. We report the evaluation results on validation set during training process for ablation experiments. The models are trained over 100 epochs with batch size 16. DML represents the decoupled multimodal learning submodule and ODA represents the offset-guided deformable alignment submodule.

of 0.002 with cosine scheduling, and the epoch is set to 50 with a batch size of 16.

In Stage II, we train the whole network with the Cross-modality Spatial Offset Modeling (CSOM) module and the Offset-guided Deformable Alignment and Fusion (ODAF) module to achieve adaptive feature alignment that is most conductive to the detection task. During this process, the $\mathcal{B}_m$ is initialized with the weights of stage I. Eventually, our OAFA is trained end-to-end with the total loss function $\mathcal{L}_{total}$. The epoch in stage II is set to 100 while other experimental parameters are set in accordance with stage I.

A.4. The Network Structure of Our OAFA

The detailed network structure information of OAFA is shown in Tab. 1.

B. Additional Ablation Experiments

B.1. Ablation Experiments on Deformable Convolutional Network

Deformable convolutional networks(DCN) [2] has inherently advantages in capturing the geometric transformation of objects, whose superior performance has been proved in object detection [14]. We conduct a series of ablation studies to verify that it is our proposed Offset-guided Deformable Alignment (ODA) module rather than the geometric modeling capability of DCN that really achieves a significant improvement in the detection task.

Since our OAFA only introduces a single DCN layer in ODA to adapt the RGB feature positions, the compared experiments are conducted exclusively in RGB modality with a DCN layer. The results are presented in Tab. 2 and Fig. 3.

Table 1. The detailed structure information of OAFA with multi-level fusion architecture. The fusion level ranges from I to III. CBL is the Convolutional Block Layer, CSP is the Cross Stage Partial Layer, SPPF is the Spatial Pyramid Pooling Fusion Layer, and DCN is the Deformable Convolutional Network Layer.

Structure		Level	Layer	Input Size	Network Type	Output Size
Unimodal Multiscale Feature Extractor $\mathcal{B}_m$		I	Focus	(640, 640, 3/1)	Conv2d, BatchNorm2d, SiLU	(320, 320, 32)
			CBL	(320, 320, 32)	Conv2d, BatchNorm2d, SiLU	(160, 160, 64)
			CSP1-1	(160, 160, 64)	CBL*3, Bottleneck	(160, 160, 64)
			CBL	(160, 160, 64)	Conv2d, BatchNorm2d, SiLU	(80, 80, 128)
			CSP1-2	(80, 80, 128)	CBL*3, Bottleneck*2	(80, 80, 128)
		II	CBL	(80, 80, 128)	Conv2d, BatchNorm2d, SiLU	(40, 40, 256)
			CSP1-3	(40, 40, 256)	CBL*3, Bottleneck*3	(40, 40, 256)
		III	CBL	(40, 40, 256)	Conv2d, BatchNorm2d, SiLU	(20, 20, 512)
			CSP2-1	(20, 20, 512)	CBL*3, Bottleneck	(20, 20, 512)
			SPPF	(20, 20, 512)	CBL*2, Maxpool*3	(20, 20, 512)
Decoupled Multimodal Learning Module	Invariant Feature Encoder $\mathcal{C}_m$	I	CBL	(80, 80, 128)	Conv2d, BatchNorm2d, SiLU	(80, 80, 128)
		II	CBL	(40, 40, 256)	Conv2d, BatchNorm2d, SiLU	(40, 40, 256)
		III	CBL	(20, 20, 512)	Conv2d, BatchNorm2d, SiLU	(20, 20, 512)
	Specific Feature Encoder $\mathcal{S}_m$	I	CBL	(80, 80, 128)	Conv2d, BatchNorm2d, SiLU	(80, 80, 128)
			CSP1-3	(80, 80, 128)	CBL*3,Bottleneck*3	(80, 80, 128)
		II	CBL	(40, 40, 256)	Conv2d, BatchNorm2d, SiLU	(40, 40, 256)
			CSP1-3	(40, 40, 256)	CBL*3,Bottleneck*3	(40, 40, 256)
		III	CBL	(20, 20, 512)	Conv2d, BatchNorm2d, SiLU	(20, 20, 512)
			CSP1-3	(20, 20, 512)	CBL*3,Bottleneck*3	(20, 20, 512)
Spatial Offset Modeling Module	Spatial Difference Enhanced	I	Spatial Attention	(80, 80, 128)	Conv2d, Sigmoid	(80, 80, 128)
		II	Spatial Attention	(40, 40, 256)	Conv2d, Sigmoid	(40, 40, 256)
		III	Spatial Attention	(20, 20, 512)	Conv2d, Sigmoid	(20, 20, 512)
	Channel Difference Enhanced	I	Channel Attention	(80, 80, 128)	Maxpool, Avgpool, RelU, FC*2, Sigmoid*2	(80, 80, 128)
		II	Channel Attention	(40, 40, 256)	Maxpool, Avgpool, RelU, FC*2, Sigmoid*2	(40, 40, 256)
		III	Channel Attention	(20, 20, 512)	Maxpool, Avgpool, RelU, FC*2, Sigmoid*2	(20, 20, 512)
Offset-guided Deformable Alignment Module		I	DCN	(80, 80, 128)	Conv2d*2, RelU*2, Sigmoid	(80, 80, 128)
		II	DCN	(40, 40, 256)	Conv2d*2, ReLU*2, Sigmoid	(40, 40, 256)
		III	DCN	(20, 20, 512)	Conv2d*2, ReLU*2, Sigmoid	(20, 20, 512)
Decoupled Feature Fusion Module		I	Fusion	(80, 80, 256)	Conv2d	(80, 80, 128)
		II	Fusion	(40, 40, 512)	Conv2d	(40, 40, 256)
		III	Fusion	(20, 20, 1024)	Conv2d	(20, 20, 512)
Detection Net		I	Prediction	(80, 80, 128)	-	(80, 80, 30)
		II	Prediction	(40, 40, 256)	-	(40, 40, 30)
		III	Prediction	(20, 20, 512)	-	(20, 20, 30)

Table 2. Ablation study on DCN. The baseline model is SLBAF-Net. DML denotes the decoupled multimodal learning submodule and ODA denotes the offset-guided deformable alignment submodule. Best results are highlighted in **bold**.

Detectors	w/ or w/o Decouple	Car	Truck	Freight-car	Bus	Van	mAP (%) ↑
Baseline	w/o Decouple	90.2	72.0	68.6	89.9	59.9	76.1
Baseline+DCN		90.2	68.1	67.5	89.9	61.4	75.4
Baseline+ODA		90.2	74.9	72.8	90.0	61.0	77.8
Baseline+DML	w/ Decouple	90.3	72.7	70.7	90.1	64.6	77.7
Baseline+DML+DCN		90.3	72.3	71.4	89.9	61.0	77.0
Baseline+DML+ODA		90.3	**75.4**	**73.7**	90.2	**65.5**	**79.0**

Table 3. Ablation study on DML loss. We exclusively remove the DML loss ($\mathcal{L}_{sim}$, $\mathcal{L}_{sim}$, and $\mathcal{L}_{sem}$) from the overall loss function in our model while preserving other components and settings. Best results are highlighted in **bold**.

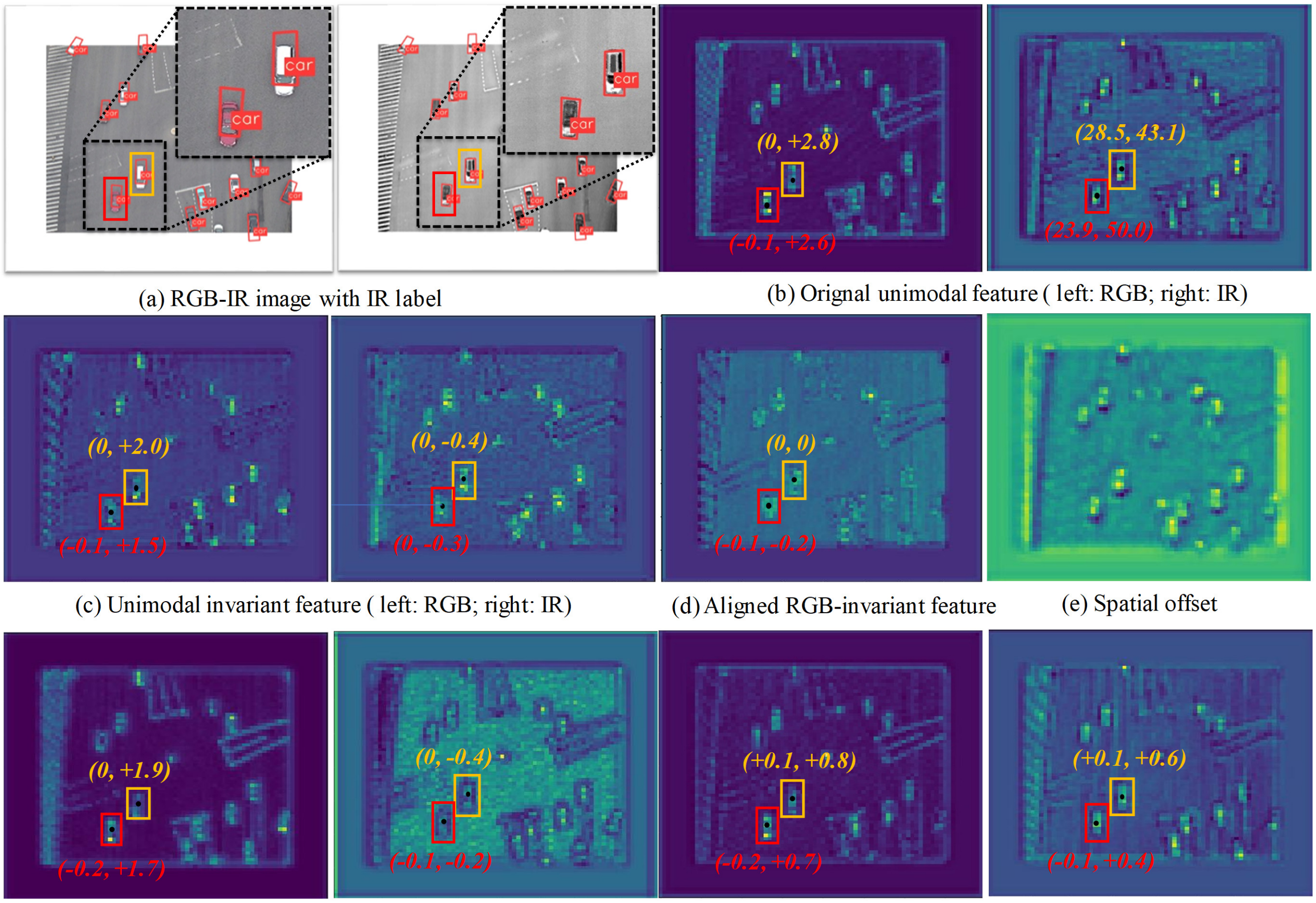
$\mathcal{L}_{sim}$	$\mathcal{L}_{dif}$	$\mathcal{L}_{sem}$	Car	Truck	Freight-car	Bus	Van	mAP (%) ↑
			90.3	74.2	70.4	90.3	64.2	77.9
✓			90.3	75.8	73.0	90.3	65.4	78.9
	✓		90.3	74.6	72.6	90.3	65.3	78.6
		✓	90.3	75.9	**73.3**	90.3	65.8	79.1
✓	✓		90.3	74.5	71.0	90.3	65.3	78.3
	✓	✓	90.3	75.3	72.6	90.2	64.4	78.5
✓		✓	90.3	75.6	73.0	90.2	64.9	78.8
✓	✓	✓	90.3	**76.8**	**73.3**	90.3	**66.0**	**79.4**

Firstly, in the baseline model, we replace a single convolutional layer in the last layer of the RGB modality feature extraction network with a DCN layer. The second row in Tab. 2 shows that the detection performance of the above model is not improved compared with the baseline model. It may be caused by the fact that the traditional DCN layer increases uncertainties in RGB representations, which is not suitable for our fusion network. In contrast, our ODA leads to an increase of 1.7%. Then, we further evaluate the performance of the model with a DCN layer in the decoupled multimodal learning (DML) submodule. Specifically, we replace the last convolutional layer of the modality-invariant feature extraction network with a DCN layer. Similar results are also observed in the model with DML. It is worth mentioning that, as shown in Fig. 3b, incorporating DCN into the modality-invariant feature extraction network causes a training collapse for the Baseline+DML+DCN model, owning to the unstable training of DCN. The above experiments demonstrate that it is adaptive feature alignment that leads to significant improvements in the detection task. However, in the car and bus category, our method fails to achieve noticeable improvement. The observed phenomenon can be attributed to the adequate number of car in the training set as well as the distinguishable representations of the bus, making the baseline model less sensitive to the weakly misalignment of these targets. The same phenomenon is also observed in Sec. B.2. For clarity, the results in these categories are not highlighted in relative tables.

B.2. Ablation Experiments on the Decoupled Multimodal Learning Loss

To verify the effectiveness of the proposed DML loss ($\mathcal{L}_{sim}$, $\mathcal{L}_{sim}$, and $\mathcal{L}_{sem}$), we conduct a series of ablation studies on the DroneVehicle dataset [10]. For comparison, we only eliminate the DML loss from the total loss function of our model and retain other components and settings.

As shown in Tab. 3, we observe that our model can achieve an improvement of 1.0% in mAP only with the similarity loss $\mathcal{L}_{sim}$. The reason may be summarized as that the $\mathcal{L}_{sim}$ can reduce the modality gap between RGB and IR features, which is beneficial to the subsequent fusion. At the same time, decoupled with the semantic loss $\mathcal{L}_{sem}$ that could bring in more semantic information also achieves satisfactory performance. Comparatively speaking, the results with difference loss $\mathcal{L}_{dif}$ is slightly lower than the model with $\mathcal{L}_{sim}$ and $\mathcal{L}_{sem}$. The underlying reasons can be inferred that despite highlighting the dissimilarity in multimodal features helps to capture complementary information, it is hard to guarantee the spatial distribution consistency of the modality-invariant features, giving rise to false estimation of the offsets. Next, the modality-specific features $\mathcal{F}_m^c$ would not have sufficient spatial information without $\mathcal{L}_{sem}$, resulting in poor evaluation re-



(a) RGB-IR image with IR label

(b) Orignal unimodal feature (left: RGB; right: IR)

(c) Unimodal invariant feature (left: RGB; right: IR)

(d) Aligned RGB-invariant feature

(e) Spatial offset

(f) Unimodal specific feature (left: RGB; right: IR)

(g) Aligned RGB-specific feature

(h) Fused feature

Figure 4. Visualization of the intermediate results in OAFA. Take the features of level I for example. The size of the features is 80×80. The concerned weakly misalignment targets are highlighted in the red and yellow boxes. The center point coordinates of the target in the original IR features are marked around the target, which is the reference value for measuring the degree of feature deviation. The marks on other features denote the relative position of its target center points to the counterpart in the original IR features.

sults. If $\mathcal{L}_{dif}$ or $\mathcal{L}_{sim}$ was removed, our model suffers from varying degrees of decline due to the inadequate feature decoupling. Finally, the model with $\mathcal{L}_{total}$ achieves the best performance, which demonstrates that the three losses can complement each other and promote the model to achieve better performance.

C. Visualization of the Intermediate Results

To demonstrate the efficacy of our method for offset alignment, we present visualizations of the intermediate feature results in Fig. 4. Take the targets in the red and yellow boxes as examples, as presented in Fig. 4a and Fig. 4b, we find from the target position of the center point that the weakly misalignment problem in images can indeed be mapped to the corresponding features. It indicates that this problem may have an impact on the subsequent fusion. Simultaneously, it can be seen from Fig. 4b and Fig. 4c that there are modal differences in the original multimodal features, and the modality-invariant features can alleviate this difference. From the second row and the third row, we can observe that with the help of the feature spatial offsets (Fig. 4e), target positions in aligned RGB features (Fig. 4d and Fig. 4g) is closer to its in IR features (the right side of Fig. 4b) than in original RGB features (the left side of Fig. 4c and Fig. 4f). The fused features (Fig. 4h) also realize target position correction, validating the advantage of our approach in addressing weakly misalignment issues.

D. Further Discussions to the Weakly Misalignment Problem

D.1. The Comparison between Different Alignment Methods

To thoroughly show the strength of our method to solve the weakly misalignment problem, we compare OAFA with some state-of-the-art multimodal alignment methods, in-

Table 4. Detection results (mAP, in %) of the image-level alignment methods and the feature-level alignment methods. Best results are highlighted in **bold**. And the second one is marked with underline.

Detectors	Level	Car	Truck	Freight-car	Bus	Van	mAP (%) ↑
HOPC [19]	Pixel-level	81.1	63.7	55	80.3	53.6	66.7
CGRP [12]		89.9	66.4	60.8	88.9	51.3	71.4
MBNet [22]		90.1	64.4	62.4	88.8	53.6	71.9
AR-CNN [20]	Feature-level	90.1	64.8	62.1	89.4	51.5	71.6
TSFADet [18]		89.9	67.9	63.7	89.8	54.0	73.1
C^2Former [17]		90.2	68.3	64.4	89.8	58.5	74.2
Ours		**90.3**	**76.8**	**73.3**	**90.3**	**66.0**	**79.4**

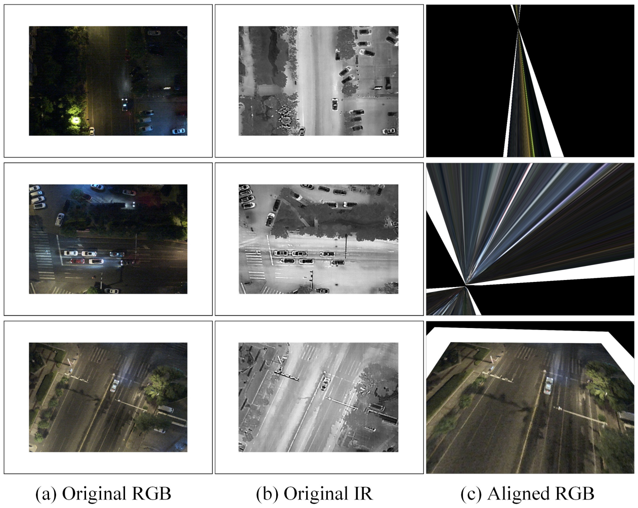

(a) Original RGB
(b) Original IR
(c) Aligned RGB

Figure 5. Three examples of error results in the image-level alignment methods.

cluding the image-level alignment methods and the feature-level alignment methods. The selected image-level alignment methods are unsupervised since the DroneVehicle dataset [10] does not provide any pixel-level alignment annotations.

The image-level alignment methods are listed as:

- HOPC [19]: This is an automatic alignment method for multimodal remote sensing data by capturing the structural similarity between images. It has been tested with a variety of optical, LiDAR, SAR, and map data.
- CGRP [12]: This method aims to accomplish unsupervised RGB-IR image alignment for the image fusion task. To achieve this goal, a robust cross-modality generation-registration paradigm is proposed to realize pixel-to-pixel multimodal image alignment.
- MBNet [22]: Based on multimodal feature learning, this method contrived a modality alignment module to predict offsets (dx, dy) for every pixel (x, y) of the sensed modality. The alignment process is constrained by the detection loss.

The feature-level alignment methods are listed as:

- AR-CNN [19]: This is the first work that tackles the weakly misalignment problem in RGB-IR object detection. AR-CNN applied an end-to-end feature alignment method in target regions to mitigate position shift.
- TSFADet [18]: This method is a further extension of AR-CNN. Considering that the targets in UAVs images are always distributed with arbitrary orientation, TSFADet not only aligned target features in positions but also in sizes and angles.
- C^2Former [17]: As an implicit alignment method, C^2Former designed an inter-modality cross-attention to obtain the calibrated features by transformer, which has the strong capability to model the pairwise correlations between multimodal features.

We adapt these methods to the RGB-IR object detection task. It is worth mentioning that for image-level alignment methods, we utilize the registered images as the input to the YOLOv5s detector. We find from Fig. 5 that the image-level alignment methods sometimes lead to error results, especially in dark night scenarios. Thus, as shown in Tab. 4, the results of the image-level alignment methods are not satisfactory. As a feature-level method, our method has excellent performance in all categories and obtains average accuracy of 79.4%, which is 5.2% higher than the second-best results. It proves that our method is more suitable for RGB-IR object detection task on UAVs under weakly misalignment conditions.

D.2. Quantitative Comparison in Robustness to Position Shift

We show the quantitative results of the position shift experiments. The scores align with the outcomes depicted in the qualitative experiments of the position shift, which has been exhibited in Sec.4.4. It can be observed from Tab. 5 that the worst result is 2.9% lower than the experiment without deviation in OAFA, whereas it increases to 4.3% in the

Table 5. Quantitative results (mAP, in %) of the position shift experiments. On the left side of the '/' are the results in baseline, and the right side are the results in our OAFA. The best results are highlighted in **bold**, while the worst results are marked with underline. The results in blue are the performance degradation compared with the best results.

mAP(%)↑		Δx						
		-15	-10	-5	0	5	10	15
Δy	-15	71.8(↓4.3%)/77.0	72.7/77.1	72.9/77.3	73.2/77.2	73.0/77.2	72.8/77.2	71.9/76.9
	-10	72.9/77.1	73.6/77.6	74.7/77.8	74.9/78.0	74.6/77.7	74.3/77.4	72.8/76.6
	-5	73.8/77.3	74.7/77.8	75.4/78.4	75.9/78.8	75.7/78.7	75.2/77.9	73.7/77.5
	0	73.6/77.3	74.8/78.1	75.8/79.0	**76.1/79.4**	75.9/79.1	75.2/78.1	73.8/77.1
	5	73.3/76.6	74.6/77.8	75.5/78.5	75.9/79.0	75.7/78.4	74.6/77.6	73.0/76.5(↓2.9%)
	10	72.8/76.5(↓2.9%)	73.5/77.2	74.5/77.7	74.7/78.6	74.4/78.3	73.4/77.2	72.5/76.6
	15	72.0/76.7	72.6/76.8	72.8/77.2	72.9/77.5	72.8/77.3	72.5/76.9	71.8(↓4.3%)/76.6

Table 6. Detection results (mAP, in %) on DroneVehicle dataset. Note that all detectors locate and classify vehicles with HBB heads. Best results are highlighted in **bold**. And the second one is marked with underline.

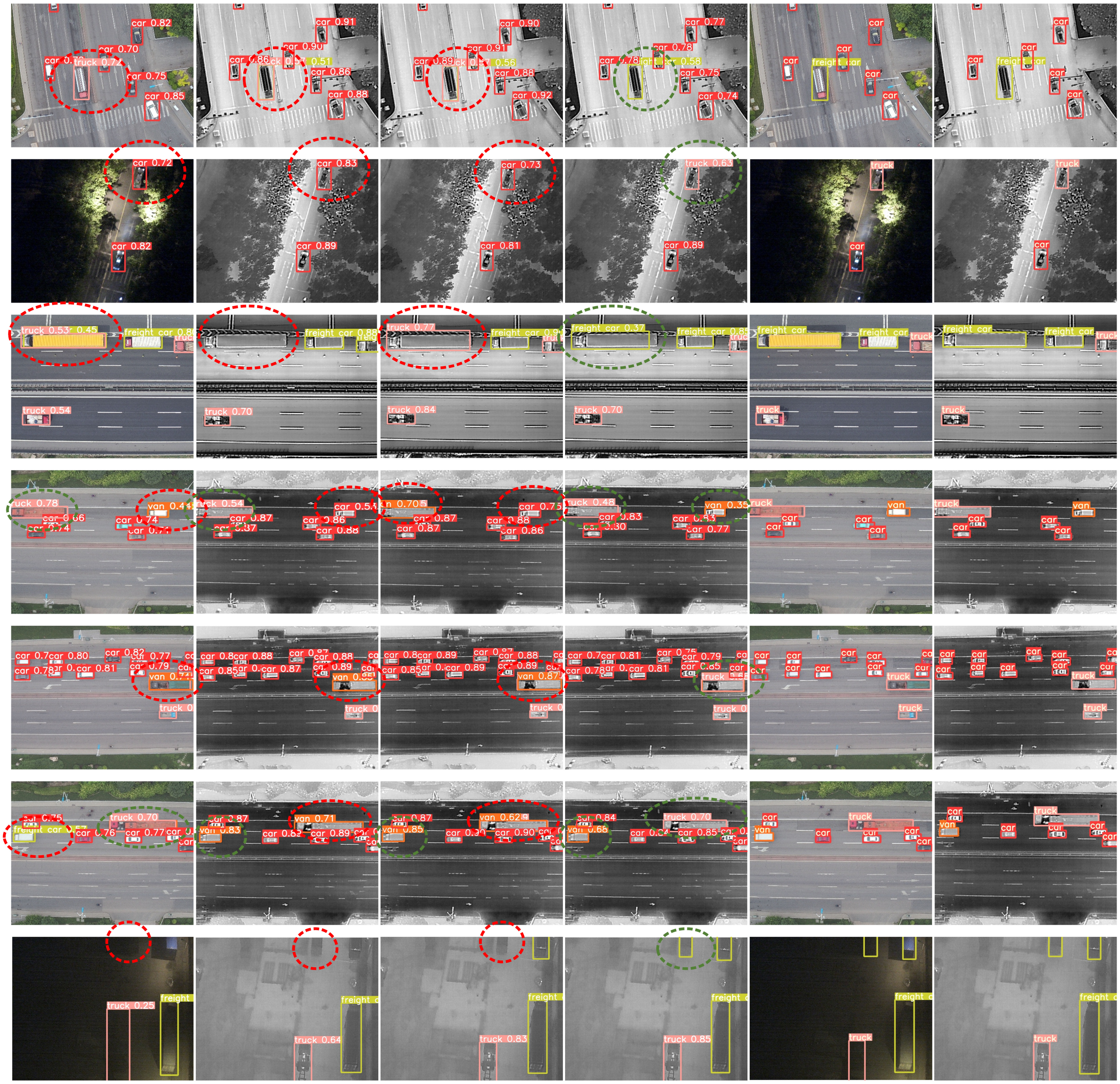
Detectors	Modality	Car	Truck	Freight-car	Bus	Van	mAP (%)↑
RetinaNet [8]	RGB	81.6	50.4	40.6	86.2	40.3	59.8
FSSD [7]		77.8	55.2	46.5	85.4	47.2	62.8
YOLOv5s [6]		91.9	70.6	56.4	92.1	56.6	73.5
YOLOX [4]		96.1	74.2	55.4	94.8	64.3	75.0
YOLOv7-Tiny* [5]		96.1	75.9	57.4	95.2	57.1	76.3
RetinaNet [8]	IR	90.3	59.3	56.6	91.4	51.3	70.6
FSSD [7]		88.5	67.5	64.6	88.7	54.4	72.7
YOLOX [4]		98.0	96.6	65.7	94.5	57.0	78.4
YOLOv5s [6]		97.6	72.5	71.5	96.1	60.7	79.6
YOLOv7-Tiny* [5]		98.1	**79.3**	67.6	95.3	58.8	79.8
PIAFusion [11]	RGB+IR	-	-	-	-	-	60.5
LAIIFusion [15]		94.5	54.4	57.9	90.5	33.9	66.2
D-ViTDet [3]		-	-	-	-	-	66.3
RISNet [13]		-	-	-	-	-	66.4
GFD-SSD [21]		90.1	71.7	61.4	89.3	57.6	74.0
AFFCM [16]		90.1	73.4	64.9	89.9	64.9	76.6
ICAFusion [9]		98.4	76.3	75.1	96.4	65.7	82.4
SLBAF-Net [1]		**98.5**	76.9	**77.3**	96.4	67.6	83.3
Ours		**98.5**	77.8	77.2	**96.8**	**69.6**	**84.0**

baseline model. The results demonstrate that our method is more robust to the weakly misalignment problem than the baseline model.

E. The Horizontal-bounding-box Experiments

E.1. Quantitative Comparison

We conduct experiments on the DroneVehicle dataset to verify the effectiveness of our method in the Horizontal-bounding-box (HBB) detection task, as many superior fusion detectors are trained with HBB annotations. To this end, we transform the original OBB annotations into HBB annotations and train the model with the same settings as the OBB experiments. We similarly compare our method with the state-of-the-art unimodal and multimodal object detectors. For unimodal detectors, we choose five outstanding methods for comparison, including RetinaNet [8], FSSD [7], YOLOv5s [6], YOLOX [4], and YOLOv7-



(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 6. Seven Examples of HBB 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.

Tiny* [5]. As for multimodal fusion detectors, our method is compared with PIAFusion [11], LAIIFusion [15], D-ViTDet [3], RISNet [13], GFD-SSD [21], AFFCM [16], ICAFusion [9], and SLBAF-Net [1]. The mAP results of the different models are shown in Tab. 6. Our OAFA achieves an mAP of 84.0%, which is 0.7% higher than the second-best method. At the same time, it also performs the best and suboptimal performances in each category. The results demonstrate that our method can also achieve excellent performance in the HBB detection task.

E.2. Qualitative Comparison

We also provide some visual HBB detection results on the validation dataset in Fig. 6. Among the compared HBB detection methods mentioned above, we select unimodal YOLOv5s [6] and the baseline model SLBAF-Net as the qualitative comparative model. It can be seen that our OAFA is prone to correct detections in different scenarios while the compared methods represent the error and missed detections. It demonstrates that the method proposed in this paper has more advantages in detection number and accuracy.

References

[1] 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

[2] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 2

[3] Zhi Fang, Tao Zhang, and XiHui Fan. A vitdet based dual-source fusion object detection method of uav. In *ICICML*, pages 628–633, 2022. 7, 8

[4] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: exceeding YOLO series in 2021. *arXiv::2107.08430*, 2021. 7

[5] Shuming Hu, Fei Zhao, Huanzhang Lu, Yingjie Deng, Jinming Du, and Xinglin Shen. Improving yolov7-tiny for infrared and visible light image object detection on drones. *Remote Sensing*, 15(13):3214, 2023. 7, 8

[6] Glenn Jocher. ultralytics/yolov5. https://github.com/ultralytics/yolov5, oct 2020. 1, 7, 9

[7] Zuoxin Li and Fuqiang Zhou. FSSD: feature fusion single shot multibox detector. *arXiv:1712.00960*, 2017. 7

[8] 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. 7

[9] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *PR*, 145: 109913, 2024. 7, 8

[10] 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. 4, 6

[11] Linfeng Tang, Jiteng Yuan, Hao Zhang, Xingyu Jiang, and Jiayi Ma. Piafusion: A progressive infrared and visible image fusion network based on illumination aware. *Information Fusion*, 83-84:79–92, 2022. 7, 8

[12] 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. 6

[13] Qingwang Wang, Yongke Chi, Tao Shen, Jian Song, Zifeng Zhang, and Yan Zhu. Improving rgb-infrared object detection by reducing cross-modality redundancy. *Remote Sensing*, 14(9):2020, 2022. 7, 8

[14] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *CVPR*, pages 14408–14419, 2023. 2

[15] Jiawen Wu, Tao Shen, Qingwang Wang, Zhimin Tao, Kai Zeng, and Jian Song. Local adaptive illumination-driven input-level fusion for infrared and visible object detection. *Remote Sensing*, 15(3):660, 2023. 7, 8

[16] Yuanfeng Wu, Xinran Guan, Boya Zhao, Li Ni, and Min Huang. Vehicle detection based on adaptive multimodal feature fusion and cross-modal vehicle index using RGB-T images. *IEEE J-STARS*, 16:8166–8177, 2023. 7, 8

[17] Maoxun Yuan and Xingxing Wei. $\mathbf{C}^2$former: Calibrated and complementary transformer for rgb-infrared object detection. *arXiv:2306.16175*, 2023. 6

[18] 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. 6

[19] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *ICCV*, pages 5126–5136, 2019. 6

[20] 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. 6

[21] Yang Zheng, Izzat H. Izzat, and Shahrzad Ziaee. GFD-SSD: gated fusion double SSD for multispectral pedestrian detection. *arXiv::1903.06999*, 2019. 7, 8

[22] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *ECCV*, pages 787–803, 2020. 6