CF-Deformable DETR: An End-to-End Alignment-Free Model for Weakly Aligned Visible-Infrared Object Detection

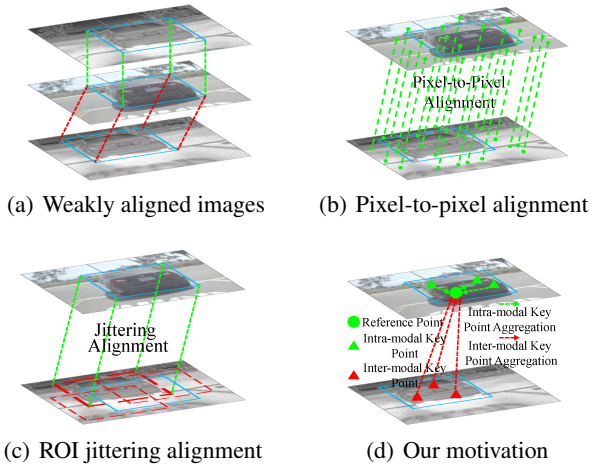
Haolong Fu, **Jin Yuan***, **Guojin Zhong**, **Xuan He**, **Jiacheng Lin** and **Zhiyong Li**†
School of Information Science and Engineering, Hunan University, Changsha, China
{haolongfu, yuanjin, gjzhong, mikasa, jcheng_lin, zhiyong.li}@hnu.edu.cn

Abstract

Weakly aligned visible-infrared object detection poses significant challenges due to the imprecise alignment between visible and infrared images. Most existing methods explore the alignment strategies between visible and infrared images, yielding unbearable computation costs. This paper first proposes an end-to-end alignment-free architecture Cross-modal Fusion Deformable DEtection TRansformer ("CF-Deformable DETR") for weakly aligned visible-infrared object detection. Abandoning the traditional image alignment, CF-Deformable DETR introduces a simple yet effective cross-modal deformable attention mechanism to directly implement automatic cross-modal point mapping, generating well-aligned bimodal features with high efficiency. Moreover, we design a Point-level Feature Consistency Loss to guide the cross-modal point mapping, ensuring the consistency of paired features to support the following fusion. Extensive experiments are conducted on three benchmark datasets. The experimental results demonstrate that CF-Deformable DETR achieves close accuracy on weakly aligned and strictly aligned data as well as maintains stable performance to a certain extent against various offset degrees of weakly aligned data. Code is available at https://github.com/116508/CF-Deformable-DETR.

1 Introduction

Object detection based on visible sensors has been widely explored and extensively applied in fields like autonomous driving [Li *et al.*, 2022b], video surveillance [Zhang *et al.*, 2022], and robotics [Feng *et al.*, 2022]. Despite the success, it is usually susceptible to a variety of environmental factors like illumination and darkness, resulting in a sharp performance drop [Li *et al.*, 2022a]. Consequently, there is a growing interest in enhancing detection performance by incorporating additional sensors, where infrared sensors demonstrate sufficient complementarity and compatibility with visible sensors, thereby promoting the increasing attention on visible-infrared object detection [Cui *et al.*, 2023].



(a) Weakly aligned images (b) Pixel-to-pixel alignment (c) ROI jittering alignment (d) Our motivation

Figure 1: An example to illustrate weakly aligned visible-infrared images (subfigure (a)) with different alignment strategies, where subfigure (b) adopts global pixel-to-pixel alignment, subfigure (c) employs local ROI jittering alignment, and subfigure (d) shows our motivation by using cross-modal alignment-free detection.

Most existing visible-infrared object detection approaches receive a strictly aligned visible (V) and infrared (IR) image pair and then effectively fuse them for performance boosting [Wang *et al.*, 2022]. Technically, there exist three main fusion strategies including pixel-level fusion [Chen *et al.*, 2022a], feature-level fusion [Yang *et al.*, 2022], and decision-level fusion [Chen *et al.*, 2022b]. No matter which paradigm is adopted, it is assumed that the input image pair is strictly aligned pixel-to-pixel as Figure 1 (a) shows. Unfortunately, in real applications, even visible and infrared sensors are fixed in a carrier and could well cooperate on image capturing, it is inevitable that the capturing V and IR images have slight pixel offsets due to the unstable moving of the carrier. As a result, the general fusion strategies usually fail on weakly aligned visible-infrared data [Zhang *et al.*, 2019].

To tackle this problem, the simple strategy first conducts a global pixel-to-pixel alignment between V and IR images (see Figure 1 (b)) before detection. This pixel-to-pixel align-

*Corresponding author1
†Corresponding author2

ment inevitably brings unbearable computation costs, thereby significantly affecting the detection speed. Another strategy incorporates the alignment process into a two-stage detector by performing a local alignment, which randomly jitters an ROI region on one modality to well match that on the other modality (see Figure 1 (c)) [Zhang *et al.*, 2019]. Despite improving efficiency, this method still cannot meet the efficiency requirement for real-world applications.

In this article, we explore whether is it possible to skip the alignment for weakly aligned visible-infrared object detection. Inspired by DETR in vision object detection as Figure 1 (d) shows, we discover that an object reference point integrating feature information from several predicted key points (see the green arrows) could offer robust point features for vision object detection. This point-based strategy leverages point features instead of region features for object detection, which offers a tip for us. If we can implement a cross-modal point-based object detection, where a reference point on one modality could automatically search and integrate relative key points from the other modality (see the red arrows), the alignment between V and IR images could be discarded.

Towards this end, this paper first proposes an end-to-end alignment-free architecture Cross-modal Fusion Deformable DEtection TRansformer ("CF-Deformable DETR") for weakly aligned visible-infrared object detection. Different from deformable DETR [Zhu *et al.*, 2020] with a single-branch decoder for single-modal object detection, CF-Deformable DETR designs a bimodal decoder with two branches, where one branch adopts single-modal key point search as same as Deformable DETR, and the other one executes cross-modal key point prediction. Concretely, given a reference point on one modality, we propose a Cross-modal Deformable Attention Mechanism (CDAM) to automatically search for relative key points on the other modality to well match the reference point. As a result, there are two features (visible and IR) generated from the two branches for each point. To guide the learning of CDAM, we design a Point-level Feature Consistency (PFC) loss, which aims to bridge the cross-modal representation gap by using contrastive learning on paired bimodal features. Differently, the PFC loss is implemented on hyperbolic space instead of Euclidean space, yielding better performance. We conduct extensive experiments on three visible-infrared object detection datasets "KAIST ", "FLIR", and "LLVIP", and the experimental results demonstrate that CF-Deformable DETR achieves promising detection accuracy with high efficiency, and could keep stable performance under varying degrees of weakly aligned data. The main contributions of this paper are summarized as follows:

- We propose an end-to-end architecture Cross-modal Fusion Deformable DEtection TRansformer for weakly aligned visible-infrared object detection. As far as our knowledge, this is the first alignment-free point-based detection network on this topic, providing a promising solution for future research.
- We propose a novel Cross-modal Deformable Attention Mechanism (CDAM) to implement cross-modal point mapping. By building a cross-modal bridging mechanism, our approach could automatically connect points between visible and IR modalities and thus avoid the cumbersome alignment with high efficiency.
- We design a Point-level Feature Consistency (PFC) loss to bridge the inter-modality representation gap, which could effectively guide the learning of CDAM, generating well-aligned visible and infrared feature pairs for performance boosting.

2 Related Work

2.1 Visible-Infrared Object Detection

The high complementarity between visible and IR representations promotes the development of visible-infrared object detection, where numerous researchers are devoted to cross-modal fusion strategies including pixel-level fusion [Cao *et al.*, 2022], feature-level fusion [Kim *et al.*, 2022; Ding *et al.*, 2021; Jain *et al.*, 2023] and decision-level fusion [Chen *et al.*, 2022b]. Among them, feature-level fusion demonstrates superior detection efficiency and effectiveness due to the deep exploration of complementary features. For example, Kinjal [Dasgupta *et al.*, 2022] enhanced the detector performance by fusing the features with contextual information extracted from four RNN networks. Liu [Liu *et al.*, 2021] introduced a deep cross-modalities representation learning-based pedestrian detection network (DCRL-PDN). Besides the aforementioned convolution-based methods, Xing [Xing *et al.*, 2023] extended the Transformer framework to multi-modal detection. By balancing and optimizing visible, infrared, and fusion detection heads, they enhanced the detector's adaptability to varying environments. This advancement brings new possibilities to the realm of visible-infrared object detection. However, it is noted that all of these efforts heavily rely on tightly aligned visible-infrared image pairs, generating huge alignment costs for real-world applications.

2.2 Weakly Aligned Visible-Infrared Object Detection

For weakly aligned visible-infrared object detection, the direct feature fusion usually suffers from a sharp performance drop. To tackle this problem, the simple solution first performs global pixel-to-pixel cross-modal alignment followed by object detection on aligned data, bringing high alignment costs [Deng and Ma, 2022]. To reduce costs, Zhang [Zhang *et al.*, 2019] integrated both alignment and detection into an end-to-end network by jittering regions of interest (ROIs), obtaining aligned ROI features to enhance detection accuracy. On this basis, Zhang [Zhang *et al.*, 2021] further extended this method to tasks involving multiple modalities like visible depth. Zhou [Zhou *et al.*, 2020] designed an alignment module that first predicts the offset for each pixel, and then obtains the final pixel value from neighboring pixels. However, the presence of information deviations has an impact on the performance of the detector. Chen [Chen *et al.*, 2022b] aggregates the detection results from multiple models on different modalities and then selects the final bounding box through post-processing, thereby avoiding the impact of weakly aligned data.

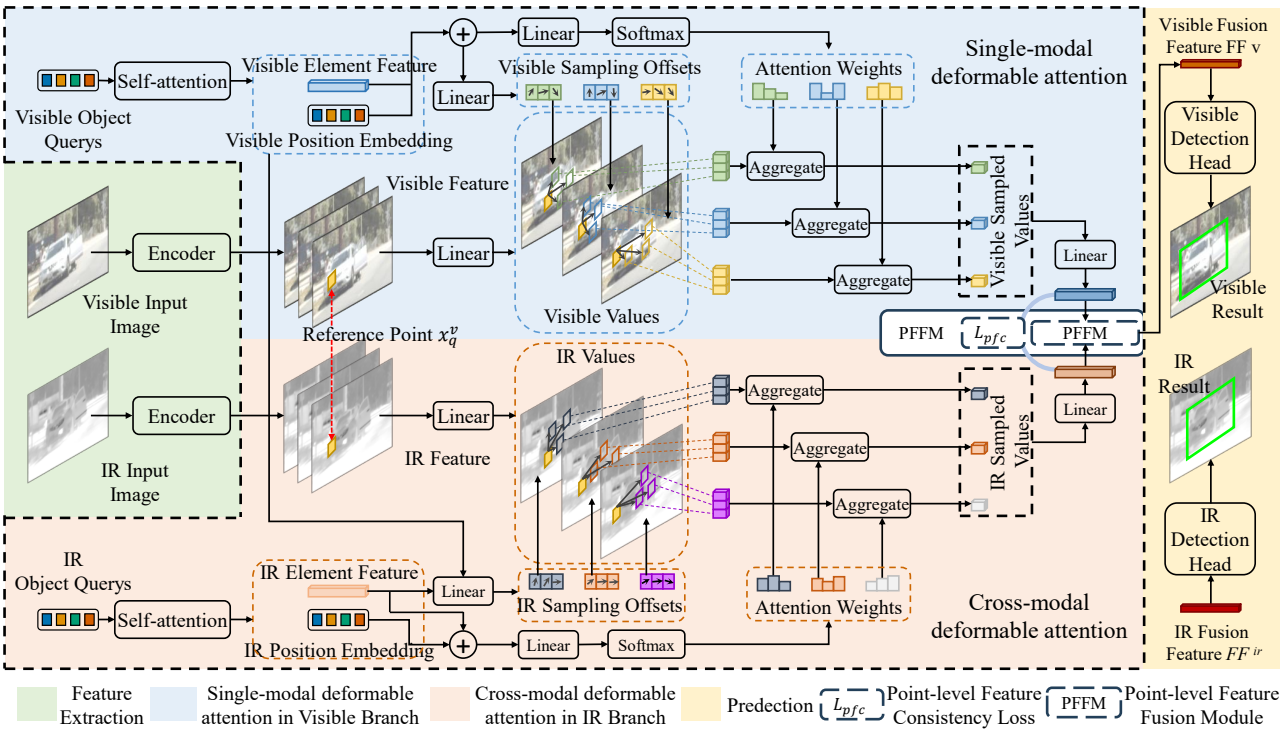


Figure 2: Overview of Cross-modal Fusion Deformable DEtection TRansformer, which consists of three parts including an encoder, a bimodal decoder, and a prediction head.

Beyond the previous studies, our work aims at completely abandoning the existing alignment strategies for weakly aligned visible-infrared object detection. Inspired by DETR, we propose an end-to-end alignment-free model CF-Deformable DETR to automatically implement cross-modal point mapping learning, supporting robust object detection with high efficiency.

3 Methodology

3.1 Architecture Overview

Figure 2 illustrates the architecture of a Cross-modal Fusion Deformable DEtection TRansformer, consisting of three parts: an encoder for extracting both visible and IR features, a bimodal decoder for generating and fusing point-level visible and IR features, and a prediction head for object prediction. Different from the single-modal Deformable DETR, CF-Deformable DETR supports receiving a weakly aligned visible and IR image pair to predict objects on both modalities. Technically, the proposed bimodal decoder contains a Single-modal Deformable Attention Module (SDAM), a Cross-modal Deformable Attention Module (CDAM), and a Bimodal Adaptive Fusion Module (BAFM). Given a reference point on a visible image, SDAM aims at predicting several relative key points to generate an aggregated visible feature which is the same as Deformable DETR. Differently, CDAM attempts to search for cross-modal key points on the IR image to generate an IR feature. We design a novel Point-level Feature Consistency (PFC) loss to guide the learning of CDAM, and the generated bimodal features are then passed to BAFM for effective feature fusion for object detection. Next, we elaborate on CDAM, PFCloss, and PFFM, respectively, followed by the implementation of our model.

3.2 Cross-Modal Deformable Attention Module

Given a reference point x_q^m ($m \in \{V, IR\}$) on the modality m with the point feature $y_q^m \in \mathbb{R}^C$ and position embedding $p_q^m \in \mathbb{R}^C$ where C is the channel dimension, cross-modal deformable attention module (CDAM) aims to search for relative key points on the other modality $\overline{m}$ to aggregate their features around x_q^m. The traditional single-modal deformable attention module predicts point offsets and attention weights on the same modality to generate an aggregated feature $f_q^m \in \mathbb{R}^C$ of x_q^m based on y_q^m and p_q^m. Differently, CDAM requires exploring the cross-modal feature correlations and thus we further add $y_q^{\overline{m}}$ to help predict point offsets and attention weights on the other modality $\overline{m}$. Specifically, our method consists of two steps: First, we concatenate the two modality features y_q^m and $y_q^{\overline{m}}$, followed by a mapping function to capture the complex correlations between them:

$$\hat{y}_q = \varphi([y_q^m, y_q^{\overline{m}}]), \tag{1}$$

where $[,]$ represents the concatenation operation at the channel level, $\varphi(\cdot)$ is a linear layer to capture the inter-modal correlations, and $\hat{y}_q \in \mathbb{R}^C$ is the generated fusion feature. Second, to generate cross-modal offsets concerning x_q^m, we incorporate the positional embedding p_q^m into the fusion feature $\hat{y}_q$ and design a cross-modal offset predictor as follows:

$$\triangle p_q^{\overline{m}} = \phi(\varphi(\hat{y}_q) + p_q^m), \tag{2}$$

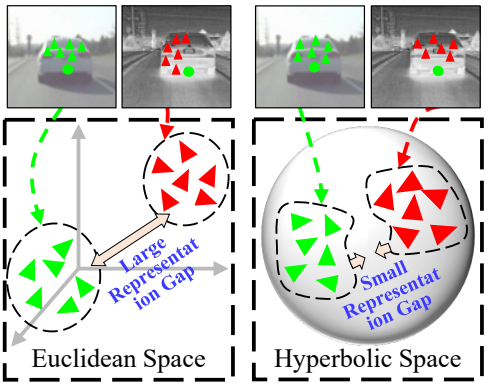


Figure 3: The motivation of Point-level Feature Consistency Loss.

where $\phi(\cdot)$ represents the re-scales operation, and $\triangle p_q^{\overline{m}}$ represents the predicted cross-modal offsets on the modality $\overline{m}$ concerning x_q^m. This cross-modal offset predictor could build a position mapping bridge between the two modalities, which automatically maps a reference point on one modality into several relative key points on the other modality, thereby effectively avoiding the pixel-level alignment problem.

Next, CDAM follows the single-modal deformable attention module to generate K key points with the corresponding attention weights by considering multiple heads, yielding an aggregated feature $f_q^{\overline{m}}$ on the modality $\overline{m}$ concerning x_q^m which is expressed as follows:

$$f_q^{\overline{m}} = \sum_{h=1}^{H} W_h^{\overline{m}} [\sum_{k=1}^{K} A_{hqk}^{\overline{m}} \cdot W_h^{'\overline{m}} y^{\overline{m}}(p_q^m + \triangle p_{qkh}^{\overline{m}})], \quad (3)$$

where $H = 8$ and $K = 4$ denote the indexes of the attention head and the key point, respectively. $W_h^{\overline{m}} \in \mathbb{R}^{C \times C_v}$ and $W_h^{'\overline{m}} \in \mathbb{R}^{C_v \times C}$ are the learnable weights ($C_v = C/H$). $A_{qhk}^{\overline{m}}$ is the scalar attention weight of the k-th key point on the h-th attention head, which is obtained via a linear projection as same as the single-modal deformable attention module. $y^{\overline{m}}(.)$ returns the key point $\overline{m}$ feature according to position information.

Injected by visible and IR feature maps, CDAM automatically builds a point mapping mechanism between the two modalities, which could generate well-aligned visible and IR features for each reference point. Therefore, our method no longer requires point-to-point image alignment operations, thereby saving amounts of computational costs.

3.3 Point-Level Feature Consistency Loss in Hyperbolic Geometry

Considering the unsupervised mapping by CDAM may suffer from the cross-modal representation gap stemming from two aspects: the inherent representation difference caused by visible and infrared modalities, and the inaccurate mapping difference as the left two images in Figure 3 shows, where the corresponding visible and infrared points focus on different components of the car. Therefore, it is necessary to bridge the cross-modal representation gap. Motivated by this, we design a Point-level Feature Consistency (PFC) loss, which aims at reducing the inherent representation difference as well as guiding the cross-modal mapping by CDAM. Since visible and infrared features inherently possess substantial representation differences, existing Euclidean-space-based metric methods cannot accurately measure the similarity between them (see Figure 3). In contrast, hyperbolic space $\mathbb{D}^n$, as a Riemannian manifold with constant negative curvature [Yaras *et al.*, 2022], gives rise to geometric properties such as curvature, symmetry, and non-linearity [Khrulkov *et al.*, 2020], making it better suited to capture domain-invariant features. Inspired by its successful application in image retrieval, we incorporate hyperbolic embedding into the proposed PFC loss, bridging the representation gap between visible-infrared feature pairs in hyperbolic geometry as well as guiding the cross-modal mapping by CDAM (see the right part in Figure 3).

Given a Euclidean vector a, the exponential bijective mapping $\exp_{\mathrm{z}}^c : T_p\mathbb{R}^n \rightarrow \mathbb{D}^n$ of the Poincaré ball modal $\mathbb{D}^n = \{\mathrm{a} \in \mathbb{R}^n | c\,\|a\|^2 < 1, c \geq 1\}$ with the curvature parameter c projects a into hyperbolic space, denoted as:

$$\exp_{\mathrm{z}}^c(\mathrm{a}) = \mathrm{z} \oplus_c (tanh(\sqrt{c}\frac{\lambda_{\mathrm{z}}^c \|\mathrm{a}\|}{2})\frac{\mathrm{a}}{\sqrt{c}\,\|\mathrm{a}\|}), \quad (4)$$

where $\oplus_c$ is the differentiable Möbius addition, z is the reference point empirically set to 0 for simplified computation, c and $\lambda_{\mathrm{z}}^c = \frac{2}{1-c\|z\|^2}$ denote the conformal factor that scales the local distance. Specifically, we first follow Eq. 4 to obtain the hyperbolic representations $\hat{f_q^m}$ and $\hat{f_q^{\overline{m}}}$ by applying the exponential mapping to f_q^m and $f_q^{\overline{m}}$ extracted from SDAM and CDAM. Then, Our PFC loss $\mathcal{L}_{pfc}$ explicitly optimizes the features of two modalities to bridge the cross-modal representation gap. Inspired by contrastive learning [Chen *et al.*, 2020], $\mathcal{L}_{pfc}$ encourages the cross-modal feature pair $\hat{f_q^m}$ and $\hat{f_q^{\overline{m}}}$ concerning the same point x_q^m to be as similar as possible, which is formulated as:

$$\mathcal{L}_{pfc} = -\frac{1}{|Q|}\sum_{q=1}^{Q} \log \sigma(\hat{f_q^m} \cdot \hat{f_q^{\overline{m}}}), \quad (5)$$

where $\sigma(\cdot)$ is the sigmoid function, and Q is the number of input points. The introduction of $\mathcal{L}_{pfc}$ has two advantages: First, it aligns the key point features from two heterogeneous modalities, reducing their cross-modal representation differences in a high-dimensional feature space. Second, it effectively guides the learning of CDAM, implementing accurate cross-modal point mapping.

3.4 Point-Level Feature Fusion Module

To better leverage the complementarity between visible and infrared features, we adopt a point-level feature fusion module to adaptively adjust the correlation weights for feature fusion. Concretely, we first construct the cross-modal correlation matrix $\mathbf{C}^m \in \mathbb{R}^{Q \times Q}$ by performing matrix multiplication and softmax function $softmax(\cdot)$ as follows:

$$\mathbf{C}^m = softmax(\mathbf{f}^{\overline{m}}\mathbf{f}^{m\mathrm{T}}) = c_{pq}^m, \quad (6)$$

	SDAM	CDAM^{-}	CDAM
V	70.7%	63.3%	**77.0%**
IR	70.9%	63.1%	**77.1%**

Table 1: Performance comparison of different cross-modal point mapping strategies tested on FLIR$_w$ dataset.

Offset		0	5	10	20	30	50
FLIR	V	**77.2%**	77.1%	77.0%	73.2%	67.5%	61.7%
	IR	**77.1%**	77.1%	77.1%	73.3%	67.7%	61.6%
LLVIP	V	**95.0%**	95.0%	94.9%	94.8%	92.2%	80.7%
	IR	**95.0%**	95.0%	95.0%	94.7%	91.9%	80.8%

Table 2: Performance comparison results on weakly aligned data with different offsets.

where $\mathbf{f}^m \in \mathbb{R}^{Q\times C}$ is a point feature matrix in the modality m and $c_{pq}^{\cdot}$ describes the inter-modality feature correlation between x_q and x_p. Similarly, we can obtain $\mathbf{C}^{\overline{m}} \in \mathbb{R}^{Q\times Q}$. By mining the cross-modal feature correlation, we can adaptively query effective information from the complementary modality when the point features of a modality cannot accurately capture an object. Then, we multiply $\mathbf{C}^m$ with $\mathbf{f}^m$, followed by performing residual calculations to obtain the correlation-weighted features $\mathbf{CF}^m$:

$$\mathbf{CF}^m = \mathbf{C}^m\mathbf{f}^m + \mathbf{f}^m, \tag{7}$$

Similarly, we obtain $\mathbf{CF}^{\overline{m}}$ in the same manner, and the two correlation-weighted features are combined as the fused feature $\mathbf{FF}^m$ by point-wise addition:

$$\mathbf{FF}^m = \mathbf{CF}^m + \mathbf{CF}^{\overline{m}}, \tag{8}$$

The fusion feature $\mathbf{FF}^m$ is obtained by integrating two heterogeneous modal features at the point level, offering richer object information for object detection.

3.5 Implementation

The above fusion module provides two fusion features $\mathbf{FF}^v$ and $\mathbf{FF}^{ir}$ to the two prediction heads to generate detection results on visible and IR images, respectively. For each branch, we adopt end-to-end training by using a composite loss function consisting of two components: the detection loss $\mathcal{L}_{det}$ and the PFC loss $\mathcal{L}_{pfc}$, denoted as:

$$\mathcal{L} = \mathcal{L}_{det} + \lambda\mathcal{L}_{pfc}, \tag{9}$$

where λ is a balancing weight. Here, $\mathcal{L}_{det}$ measures the deviation between ground truth and prediction results. More details of $\mathcal{L}_{det}$ could be referred to [Zhu *et al.*, 2020].

4 Experiments

4.1 Datasets

We adopt three datasets KAIST [Hwang *et al.*, 2015], FLIR [FLIR, 2018] and LLVIP [Jia *et al.*, 2021] as the baseline to construct three weakly aligned datasets as follows:

- **KAIST**: This dataset is a challenging dataset for pedestrian detection. We adopt the improved dataset by [Liu *et al.*, 2016] containing $7,601$ pairs for training and $2,252$ pairs for testing.

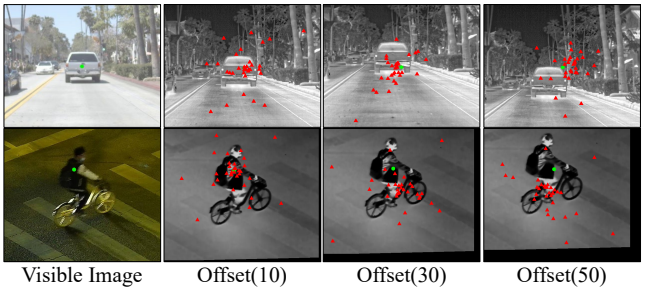


Figure 4: Two examples illustrate the cross-modal attention results, where the right three images have different offsets. The green cycle represents a reference point, and the red triangles are the generated 32 ($K = 4$, $H = 8$) cross-modal key points.

- **FLIR**: We use a cleaned version in [Zhang *et al.*, 2020] including $4,129$ and $1,013$ visible/IR image pairs for training and testing with the image size 640×512.
- **LLVIP**: This dataset is captured in low-light environments for pedestrian detection. There are a total of $15,488$ visible/IR image pairs (1280×1024) divided into $12,025$ pairs for training and $3,463$ for testing.
- **Weakly Aligned Dataset Generation**: We create weakly aligned image pairs (KAIST$_w$", FLIR$_w$," "LLVIP$_w$") by introducing random offsets on IR images while fixing visible images. Specifically, we use the horizontal and vertical offset with $[-10, 10]$ pixels for FLIR, and $[-20, 20]$ pixels for LLVIP as considering their resolution difference. Differently, we construct the KAIST$_w$ dataset following [Zhang *et al.*, 2019], which adds two more offset directions including $45°$ and $135°$.

4.2 Experimental Setups

All the experiments are deployed on NVIDIA RTX A6000 GPUs. Specifically, we adopt Resnet50 as the backbone and the transformer encoder with the same configuration of Deformable DETR. For the decoder, we set the number of reference points Q to 300. The weight λ of $\mathcal{L}_{pfc}$ is set to 0.05. To train our model, we execute a total of 60 epochs with the batch size 4 for all the datasets by using the Adam optimizer. The initial learning rate is 0.0001 and decayed by a factor of 0.1 at the 40-th epoch. All the testing results are reported by mAP at $IOU = 0.5$ (mAP@50), which is widely used in visible-infrared object detection [Chen *et al.*, 2022b].

4.3 Evaluation on CDAM

We compare CDAM with two different cross-modal point mapping mechanisms as shown in Table 1, where SDAM directly uses the coordinates of key points in the single-modal branch to select key points in the cross-modal branch, and CDAM^{-} aims at mapping a reference point in the single-modal branch to a corresponding point in the cross-modal branch. Compared to CDAM, SDAM declines about 6% mAPs since selecting the two cross-modal points with the same coordinate for object detection cannot overcome the pixel offset problem on weakly aligned data, resulting in unpaired visible and IR feature generation. Moreover, CDAM^{-}

(a) KAIST$_w$ (b) FLIR$_w$ (c) LLVIP$_w$

Figure 5: Three examples illustrate the detection results by different approaches, where the orange triangles indicate false detection.

$\mathcal{L}_{pfc}$ in $\mathbb{R}^n$	$\mathcal{L}_{pfc}$ in $\mathbb{D}^n$	mAP	
		V	IR
✓	×	76.4%	76.5%
×	✓	**77.0%**	**77.1%**

Table 3: Performance comparison by using or not using $\mathcal{L}_{pfc}$ tested on FLIR$_w$ dataset.

λ	0	0.01	0.05	0.1	0.5	1
V	75.1%	75.9%	**77.0%**	75.8%	75.4%	75.2%
IR	75.0%	75.8%	**77.1%**	75.9%	75.3%	75.2%

Table 4: Performance trend of CF-Deformable-DETR with different λ in Equation 9 tested on FLIR$_w$ dataset.

achieves the worst performance because it adopts a two-stage point mapping including searching for a cross-modal reference point followed by searching for cross-modal key points, resulting in the error accumulation problem in point selection.

We further generate weakly aligned datasets with different offsets to verify the effectiveness of CDAM as shown in Table 2. On FLIR$_w$ dataset, the performance remains stable when the offset is less than 10 pixels. As we further increase the offset, the performance suffers from a drop but still exceeds 60% mAP values. Comparatively, LLVIP has a larger resolution to remain stable performance when the offset is less than 20 pixels, which proves the effectiveness of CDAM on weakly aligned data. However, we also discovered too large offsets still have a negative effect due to the increasing difficulty of cross-modal key point search. Figure 4 lists two examples to visualize the generated cross-modal key points concerning a reference point. It is shown that the corresponding reference points on the weakly aligned visible images have a certain positional deviation from that on the IR image. In such cases, CDAM is still able to identify relevant key points. Moreover, as the offset increases, the quality of key points has deteriorated with more points deviating from the objects, which is the reason for the performance degradation in Table 2.

4.4 Evaluation on PFC Loss

This experiment explores the effectiveness of the Point-level Feature Consistency loss $\mathcal{L}_{pfc}$ in Equation 5. As listed in Table 3, we compare the performance of $\mathcal{L}_{pfc}$ in Euclidean space with that in hyperbolic space. Obviously, using the loss in hyperbolic space achieves better performance, with 0.6% mAP improvement. This is because hyperbolic space effectively reduces the representation gap across modalities, thereby better guiding the learning of CDAM to offer accurate bimodal feature pairs for object detection. We further test the effect of different weights of $\mathcal{L}_{pfc}$ in Equation 9 as shown in Table 4. The best performance arrives at $\lambda = 0.05$ with mAPs of 77.0% and 77.1% on visible and IR images, respectively. The smaller values of λ overshadow the utility of $\mathcal{L}_{pfc}$, which is not conducive to feature alignment. Comparatively, too large values of λ affect the detection loss $\mathcal{L}_{det}$, resulting in performance decline.

4.5 Comparison With State-of-the-Art Methods

We compare CF-Deformable DETR with several state-of-the-art methods on KAIST, FLIR, and LLVIP datasets, and the corresponding results are shown in Tables 5, 6, and 7, respectively, where the first row part refers to strictly aligned detection methods, and the second part is weakly aligned detection methods. All the results in Table 5 are cited from the

Models	MR^{-2}				
	O	S^{0°	S^{45°	S^{90°	S^{135°
Halfway Fusion [Zhang *et al.*, 2020]	25.51%	33.73%	36.25%	28.30%	36.71%
MLPD [Kim *et al.*, 2021]	7.94%	13.34%	14.79%	8.79%	15.01%
LG-FAPF [Cao *et al.*, 2022]	**5.76%**	8.57%	10.23%	8.55%	9.67%
TFNet [Chu *et al.*, 2023]	8.09%	9.78%	10.11%	9.65%	10.17%
AANet [Chen *et al.*, 2023]	6.11%	9.13%	9.87%	9.15%	9.73%
ARCNN [Zhang *et al.*, 2019]	8.26%	9.34%	9.73%	8.91%	9.79%
MBNet [Zhou *et al.*, 2020]	9.04%	14.85%	17.62%	10.32%	17.17%
ProbEN3 [Chen *et al.*, 2022b]	7.66%	7.83%	8.17%	7.57%	8.32%
CF-Deformable DETR	V: 7.13%	**7.14%**	**7.17%**	**7.14%**	**7.17%**
	IR: 7.11%	**7.14%**	**7.15%**	**7.15%**	**7.17%**

Table 5: Performance comparison between our approach and the advanced methods on KAIST$_w$ datasets measured by MR^{-2}. Following the settings in ARCNN, we test these modules along 4 directions (S^{0°, S^{45°, S^{90° and S^{135°), and O denotes no position shift.

Models	mAP
	FLIR $\rightarrow$ FLIR$_w$
Halfway-FS [Zhang *et al.*, 2020]	71.2% $\rightarrow$ 67.1%
Yolo-FS [Qingyun and Zhaokui, 2022]	76.6% $\rightarrow$ 67.8%
CFT [Qingyun *et al.*, 2021]	77.7% $\rightarrow$ 68.3%
LRAF-Net [Fu *et al.*, 2023]	**80.5%** $\rightarrow$ 69.9%
MBNet [Zhou *et al.*, 2020]	72.4% $\rightarrow$ 69.3%
ProbEn3 [Chen *et al.*, 2022b]	75.5% $\rightarrow$ 72.8%
CF-Deformable DETR	V:77.2% $\rightarrow$ **77.0%**
	IR:77.1% $\rightarrow$ **77.1%**

Table 6: Performance comparison between our approach and the state-of-the-art methods on FLIR and FLIR$_w$ datasets.

Models	mAP
	LLVIP $\rightarrow$ LLVIP$_w$
Halfway Fusion [Zhang *et al.*, 2020]	95.1% $\rightarrow$ 92.5%
Yolo-FS [Qingyun and Zhaokui, 2022]	95.4% $\rightarrow$ 92.7%
CFT [Qingyun *et al.*, 2021]	97.5% $\rightarrow$ 93.0%
LRAF-Net [Fu *et al.*, 2023]	**97.9%** $\rightarrow$ 93.2%
MBNet [Zhou *et al.*, 2020]	94.8% $\rightarrow$ 93.1%
ProbEn3 [Chen *et al.*, 2022b]	95.5% $\rightarrow$ 93.7%
CF-Deformable DETR	V:95.0% $\rightarrow$ **94.8%**
	IR:95.0% $\rightarrow$ **94.7%**

Table 7: Performance comparison between our approach and the state-of-the-art methods on LLVIP and LLVIP$_w$ datasets.

corresponding references, and those in Tables 6, and 7 are reproduced by using published codes. Overall, the strictly aligned detection methods suffer from a significant performance drop on weakly aligned data since they adopt weakly aligned bimodal features for fusion, thereby degrading the performance. Comparatively, weakly aligned detection methods alleviate this performance drop since they perform cross-modal image alignment before detection. Among all the weakly aligned methods, Our method performs the best and hardly leads to any performance degradation. This superior performance stems from the effective cross-modal deformable attention mechanism incorporating the feature consistency loss, which could implement accurate feature alignment to support weakly aligned object detection. Table 8 further lists the testing times of weakly aligned methods. Although our approach employs the transformer-based structure, the testing time is quite good as compared to the other CNN-based methods since our approach does not conduct cross-modal image alignment, saving a number of computational costs.

Figure 5 shows the detection results by different methods on three examples. Our approach could well detect all the objects on the two modalities, while MBNet and ProbEN3 miss several partially occluded objects. This result reflects the advantages of our cross-modal point-based detection on weakly aligned data, which explores multi-modal point features to better distinguish overlapping objects.

Detection Model	time(ms/img)
ARCNN	120.0
MBNet	**70.9**
ProbEN3	846.9
CF-Deformable DETR	79.7

Table 8: Testing time comparison among several weakly aligned approaches tested on the FLIR dataset.

5 Conclusions

In this work, we delve into weakly aligned visible-infrared object detection and first propose an end-to-end alignment-free model CF-Deformable DETR. Abandoning the existing alignment strategies, CF-Deformable DETR directly implements cross-modal point feature extraction and fusion conditioned on weakly aligned data, eliminating the alignment overhead. Technically, CF-Deformable DETR proposes a cross-modal deformable attention mechanism to automatically search for matching points as well as a point-level feature consistency loss to effectively guide the learning, offering well-aligned bimodal features for detection. Extensive experiments on three benchmark datasets confirm the effectiveness and efficiency of the proposed approach, yielding promising and stable performance on weakly aligned data. We hope our work can provide a new paradigm for solving weakly aligned object detection, promoting technological progress in this issue.

Acknowledgments

This work was partially supported by National Natural Science Foundation of China (No.U21A20518, No.U23A20341, No.62272157).

References

[Cao *et al.*, 2022] Yanpeng Cao, Xing Luo, Jiangxin Yang, Yanlong Cao, and Michael Ying Yang. Locality guided cross-modal feature aggregation and pixel-level fusion for multispectral pedestrian detection. *Information Fusion*, 88:1–11, 2022.

[Chen *et al.*, 2020] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*, pages 1597–1607. PMLR, 2020.

[Chen *et al.*, 2022a] Xu Chen, Lei Liu, and Xin Tan. Robust pedestrian detection based on multi-spectral image fusion and convolutional neural networks. *Electronics*, 11(1):1, 2022.

[Chen *et al.*, 2022b] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *European Conference on Computer Vision*, pages 139–158. Springer, 2022.

[Chen *et al.*, 2023] Nuo Chen, Jin Xie, Jing Nie, Jiale Cao, Zhuang Shao, and Yanwei Pang. Attentive alignment network for multispectral pedestrian detection. In *Proceedings of the ACM International Conference on Multimedia*, pages 3787–3795, 2023.

[Chu *et al.*, 2023] Fuchen Chu, Jiale Cao, Zhanjie Song, Zhuang Shao, Yanwei Pang, and Xuelong Li. Toward generalizable multispectral pedestrian detection. *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[Cui *et al.*, 2023] Chenhang Cui, Jinyu Xie, and Yechenhao Yang. Bright channel prior attention for multispectral pedestrian detection. *arXiv preprint arXiv:2305.12845*, 2023.

[Dasgupta *et al.*, 2022] Kinjal Dasgupta, Arindam Das, Sudip Das, Ujjwal Bhattacharya, and Senthil Yogamani. Spatio-contextual deep network-based multimodal pedestrian detection for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 23(9):15940–15950, 2022.

[Deng and Ma, 2022] Yuxin Deng and Jiayi Ma. Redfeat: Recoupling detection and description for multimodal feature learning. *IEEE Transactions on Image Processing*, 32:591–602, 2022.

[Ding *et al.*, 2021] Lu Ding, Yong Wang, Robert Laganiere, Dan Huang, Xinbin Luo, and Huanlong Zhang. A robust and fast multispectral pedestrian detection deep network. *Knowledge-Based Systems*, 227:106990, 2021.

[Feng *et al.*, 2022] Xiaoxu Feng, Xiwen Yao, Gong Cheng, and Junwei Han. Weakly supervised rotation-invariant aerial object detection network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14146–14155, 2022.

[FLIR, 2018] Teledyne FLIR. Flir thermal dataset for algorithm training. https://www.flir.in/oem/adas/adas-dataset-form, 2018. Accessed: 2019-08-16.

[Fu *et al.*, 2023] Haolong Fu, Shixun Wang, Puhong Duan, Changyan Xiao, Renwei Dian, Shutao Li, and Zhiyong Li. Lraf-net: Long-range attention fusion network for visible–infrared object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Hwang *et al.*, 2015] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection: Benchmark dataset and baseline. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1037–1045, 2015.

[Jain *et al.*, 2023] Deepak Kumar Jain, Xudong Zhao, Germán González-Almagro, Chenquan Gan, and Ketan Kotecha. Multimodal pedestrian detection using metaheuristics with deep convolutional neural network in crowded scenes. *Information Fusion*, 95:401–414, 2023.

[Jia *et al.*, 2021] Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3496–3504, 2021.

[Khrulkov *et al.*, 2020] Valentin Khrulkov, Leyla Mirvakhabova, Evgeniya Ustinova, Ivan Oseledets, and Victor Lempitsky. Hyperbolic image embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6418–6428, 2020.

[Kim *et al.*, 2021] Jiwon Kim, Hyeongjun Kim, Taejoo Kim, Namil Kim, and Yukyung Choi. Mlpd: Multi-label pedestrian detector in multispectral domain. *IEEE Robotics and Automation Letters*, 6(4):7846–7853, 2021.

[Kim *et al.*, 2022] Jung Uk Kim, Sungjune Park, and Yong Man Ro. Towards versatile pedestrian detector with multisensory-matching and multispectral recalling memory. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 1157–1165, 2022.

[Li *et al.*, 2022a] Qing Li, Changqing Zhang, Qinghua Hu, Huazhu Fu, and Pengfei Zhu. Confidence-aware fusion using dempster-shafer theory for multispectral pedestrian detection. *IEEE Transactions on Multimedia*, 2022.

[Li *et al.*, 2022b] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022.

[Liu *et al.*, 2016] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N Metaxas. Multispectral deep neural networks for pedestrian detection. *arXiv preprint arXiv:1611.02644*, 2016.

[Liu *et al.*, 2021] Tianshan Liu, Kin-Man Lam, Rui Zhao, and Guoping Qiu. Deep cross-modal representation learning and distillation for illumination-invariant pedestrian detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(1):315–329, 2021.

[Qingyun and Zhaokui, 2022] Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, page 108786, 2022.

[Qingyun *et al.*, 2021] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection. *arXiv preprint arXiv:2111.00273*, 2021.

[Wang *et al.*, 2022] Qingwang Wang, Yongke Chi, Tao Shen, Jian Song, Zifeng Zhang, and Yan Zhu. Improving rgb-infrared pedestrian detection by reducing cross-modality redundancy. In *IEEE International Conference on Image Processing*, pages 526–530, 2022.

[Xing *et al.*, 2023] Yinghui Xing, Song Wang, Guoqiang Liang, Qingyi Li, Xiuwei Zhang, Shizhou Zhang, and Yanning Zhang. Multispectral pedestrian detection via reference box constrained cross attention and modality balanced optimization. *arXiv preprint arXiv:2302.00290*, 2023.

[Yang *et al.*, 2022] Xiaoxiao Yang, Yeqiang Qian, Huijie Zhu, Chunxiang Wang, and Ming Yang. Baanet: Learning bi-directional adaptive attention gates for multispectral pedestrian detection. In *2022 International Conference on Robotics and Automation*, pages 2920–2926. IEEE, 2022.

[Yaras *et al.*, 2022] Can Yaras, Peng Wang, Zhihui Zhu, Laura Balzano, and Qing Qu. Neural collapse with normalized features: A geometric analysis over the riemannian manifold. *Advances in Neural Information Processing Systems*, 35:11547–11560, 2022.

[Zhang *et al.*, 2019] Lu Zhang, Xiangyu Zhu, Xiangyu Chen, Xu Yang, Zhen Lei, and Zhiyong Liu. Weakly aligned cross-modal learning for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5127–5137, 2019.

[Zhang *et al.*, 2020] Heng Zhang, Elisa Fromont, Sébastien Lefèvre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *IEEE International Conference on Image Processing*, pages 276–280, 2020.

[Zhang *et al.*, 2021] Lu Zhang, Zhiyong Liu, Xiangyu Zhu, Zhan Song, Xu Yang, Zhen Lei, and Hong Qiao. Weakly aligned feature fusion for multimodal object detection. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

[Zhang *et al.*, 2022] Pengyu Zhang, Jie Zhao, Dong Wang, Huchuan Lu, and Xiang Ruan. Visible-thermal uav tracking: A large-scale benchmark and new baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8886–8895, 2022.

[Zhou *et al.*, 2020] Kailai Zhou, Linsen Chen, and Xun Cao. Improving multispectral pedestrian detection by addressing modality imbalance problems. In *European Conference on Computer Vision*, pages 787–803. Springer, 2020.

[Zhu *et al.*, 2020] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020.