CAFF-DINO: Multi-spectral object detection transformers with cross-attention features fusion

Kevin Helvig   Baptiste Abeloos   Pauline Trouvé-Peloux

DTIS, ONERA, Université Paris-Saclay, 91120 Palaiseau, France
{kevin.helvig, baptiste.abeloos, pauline.trouve}@onera.fr

Abstract

Object detection on images can find benefit from coupling multiple spectra, each presenting specific useful features. However, building an efficient architecture coupling the different modalities is a complex task. Transformers, due to their ability to extract meaningful correlations between the different regions of the inputs appear as a promising way to perform features fusion across different spectra. This work presents a multi-spectral object detection architecture based on cross-attention features fusion (CAFF), combined with a transformer based detector (DINO). We demonstrate here the performance of the proposed approach in object detection compared with state-of-the-art approaches, on infrared-visible multi-spectral datasets. Moreover the robustness to systematic misalignment between image pairs is studied. The proposed approach is generic to any mono-spectrum transformer based detectors. The model developed in this study will be available in a dedicated github repository.

1. Introduction

Visible spectrum object detection can suffer lacks of information, due to environment change thought time (day versus night for example), inducing missed detection or false alarms as illustrated in Figure 1. Multi-spectral information fusion, combining complementary information from various modalities, is important to improve object detection in such challenging situations. Therefore multi-spectral fusion is studied in the literature in many application fields such as surveillance, remote sensing or robotic vision. Several efficient deep learning models based on convolutional neural networks adapted to multi-spectral infrared (IR) and visible fusion brought strong performance in object detection by combining both of these modalities on well aligned image pairs [1].

The emergence of transformers using the attention mechanism to extract associations between the different regions of the input, opened new possibilities for fusion [2]. In the context of text-image multi-modal fusion, the cross-attention operation has demonstrated great capabilities in text-image fusion [3, 4]. However, to the best of our knowledge, this operation hasn't been a lot investigated in the literature of IR-visible image pairs fusion. Besides, if transformer based detectors are now used in many applications in computer vision, their benefit for the task of image fusion has not yet been studied. Finally, image acquisition can be prone to alignment errors and robustness to imperfect alignment of both images of the pairs is not explored in the literature.

In this paper we propose a new IR-visible object detection transformer based on features fusion method (called CAFF) using cross-attention mechanism, and a modern transformer-based detector (DINO). We show that this model outperforms the state-of-the-art approaches on several public datasets. The proposed model is generic, able to be rapidly implemented on most of mono-modal transformer-based detectors of the literature. We also study the robustness of CAFF-DINO to systematic misalignment between the image pairs.

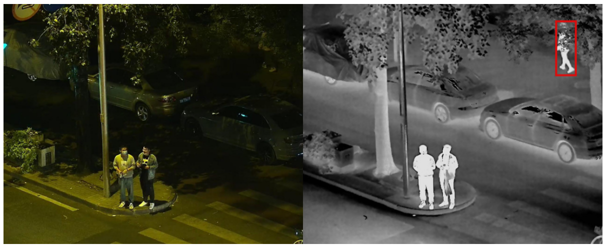

Figure 1. From left to right: visible and IR image patches from LLVIP dataset in low-light situation [5]. The reduced enlightenment illustrates the benefit of IR: the third pedestrian, framed in red, is barely noticeable in the visible spectra.

2. Related work

Multi-spectral image fusion using deep learning is an open research field, with paired datasets in open access since several years [1]. One of the first work deploying convolutional neural networks to fuse multi-spectral data was proposed in [6]. Following works have focused their studies on where to fuse features from both modalities in order to give the most important performance increase, using concatenation or other standard matrix operations [7]. Especially, following the terminology used in [8],early, mid-fusion and late fusion, have been compared in [9], considering similar architectures. It is generally admitted that mid-fusion and late fusion generally gives better localization performances, whereas earlier feature fusion struggles more to fuse efficiently features[8, 9]. However later fusers can become easily computationally expensive, needing generally the equivalent of two complete models for each modality. Features fusion (mid-fusion) is lighter, but it requires to guide the general learning behavior of the model to meaningfully associate both input spectra. **Cyclic fuse-and-refine approach** is one of the first approach developed to fuse representations extracted at backbone-level: a convolution-based fused-and-refine module enriches features extracted by the model at several levels of abstraction [10]. Several approaches, engineering correlated features extraction between both spectra were developed, as through the **concatenation of modality-wise channels** or the **shuffling between mono-modal features** channels and spatial patches [8]. Several features fusion using **self-attention or comparable operation** applied at the level of the features extraction has been developed in the literature, performing backbone features fusion based on cyclic-fused and refined paradigm [2, 11]. However, these methods have been design for a specific detector head and would require adaptation to new detection module.

Transformer based detectors are a growing approach in the field of computer vision: theses architectures are able to extract meaningful correlations between regions of the input image, at various scales [12]. The great ability of transformers to process rich information at multiple scales gives the opportunity to build simpler and monolithic deep neural models for object detection and localization, avoiding the need for intermediate structures such as region proposal sub-networks. The earlier detection-localization model based on transformers and available in the literature is **DETR (DEtection TRansformer)**, which consists in a monolithic transformer encoder-decoder head based on the attention mechanism [13]. The architecture is built upon a features extractor, then attention mechanism applied on features maps to perform object detections. **Deformable-DETR** increases performance of the original DETR model by using pyramidal features extraction and deformable-attention instead of conventional attention [14]. Deformable attention uses flexible kernels instead of static ones, strengthening the capability of the model to handle various object sizes, whereas it was a limitation of the original DETR [15]. Finally, **DINO** model introduces a denoising process between encoder object queries and decoder queries, much stronger data augmentations and denoising self-supervised learning approach [16]. This model is one of the state-of-the-art detection transformers available in the very recent detection transformers literature, followed by several challengers [17, 18]. To the best of our knowledge, these detectors have not yet been widely studied in the literature for IR-visible fusion, and particularly in combination with cross-attention features fusion.

3. Paper organization

The present paper proposes a new architecture performing backbone features fusion using cross-attention operation and generic to any detection transformers models (Section 4). We show that the proposed features fusion outperforms state-of-the-art object detection performance on standard visible-IR fusion datasets (Section 5.3). To evaluate the robustness to error in image pairs registration, we perform an experiment where image pairs are experimentally misaligned. Localization performance is evaluated for various misalignment (consisting in horizontal-vertical systematic translations of the infrared image), comparing the robustness of the proposed approach with the CFT-YOLO-v5 method (Section 5.4). An ablation study is conducted in Section 5.5, to compare several transformers detectors and features extractors from the literature. The fusion method proposed is also challenged by alternative fusion operations.

4. CAFF-DINO architecture

A crucial point in the IR-visible fusion is how to efficiently combine the representations from both modalities. The architecture proposed in this work is illustrated in Figure 2. The solution extracts a new fused features map between the two mono-modal backbones features, at each level of abstraction before injection into the encoder-decoder of a transformer based detector. Compared to several features fusions based on cyclic-fused and refine approach [2, 10], the fused features extracted are not re-injected into the mono-modal backbones, isolating each features extractor from the other. This choice, and the absence of modification of the mono-modal features extractors facilitates the deployment of the model through the direct use of pre-trained weights available online. It also facilitate the interchangeability of the modalities. The detection head exploiting fused features in our proposal is the DINO detector.

The idea behind the proposed fusion is to force the meaningful association and correlations extraction between input

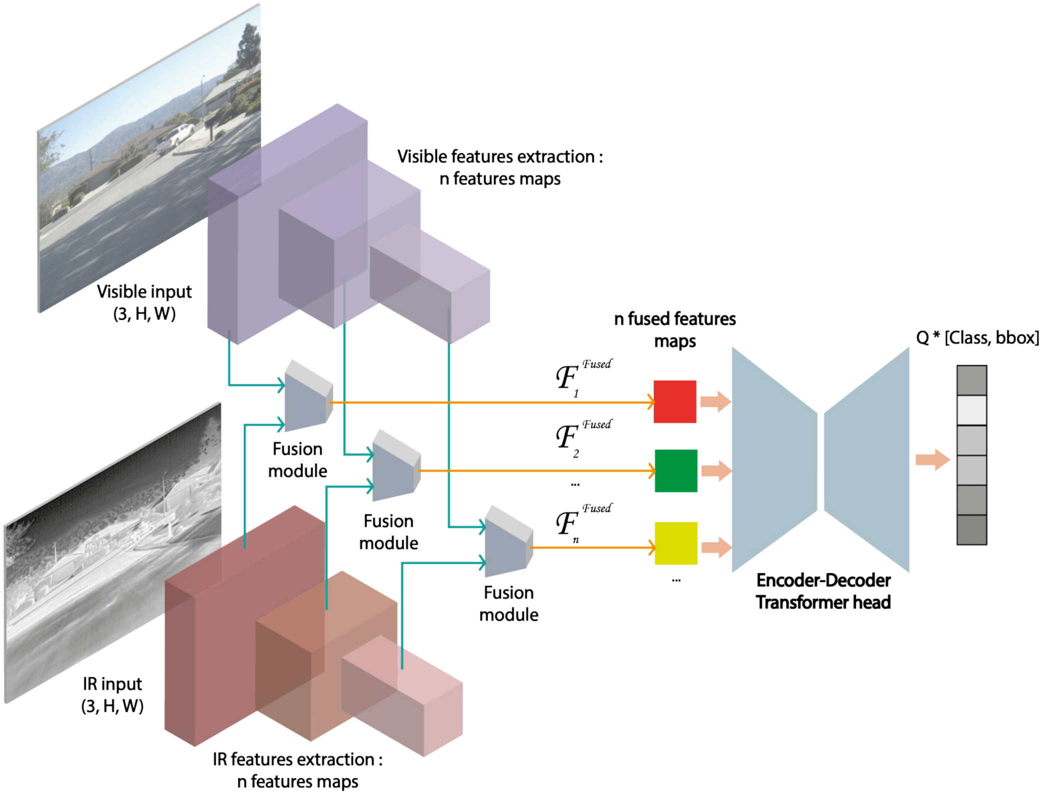

Figure 2. Illustration of the multi-spectral fusion architecture. The architecture is composed of two mono-spectral features extractor, combined with several additional fusion modules, and a transformer head that performs the object detection from both fused modalities.

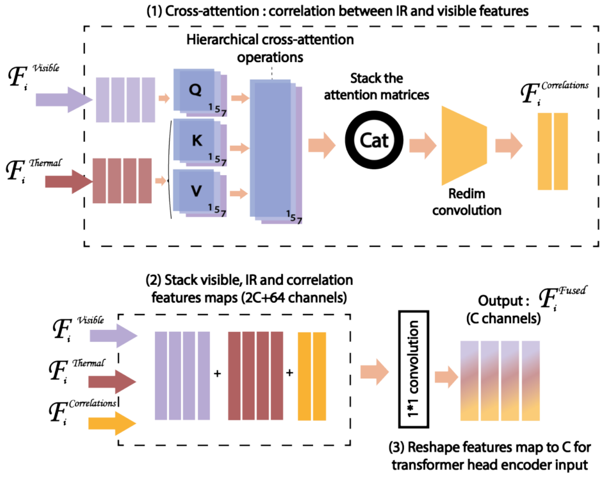
(1) Cross-attention : correlation between IR and visible features
Hierarchical cross-attention
operations
$\mathcal{F}_i^{Visible}$
$\mathcal{F}_i^{Thermal}$
Q
K
V
1 5 7
Stack the
attention matrices
Cat
Redim
convolution
$\mathcal{F}_i^{Correlations}$
(2) Stack visible, IR and correlation
features maps (2C+64 channels)
$\mathcal{F}_i^{Visible}$
$\mathcal{F}_i^{Thermal}$
$\mathcal{F}_i^{Correlations}$
+
+
1*1 convolution
Output : $\mathcal{F}_i^{Fused}$
(C channels)
(3) Reshape features map to C for
transformer head encoder input

Figure 3. Illustration of the Caff features fusion module. At each extraction level, the mono-spectral features maps are correlated using hierarchical cross-attention operations (1). The IR, Visible, and correlation features maps are then concatenated (2). These features are finally fused with 1*1 convolution operations and adapted to the detection head (3).

modalities. Instead of a previous work using self-attention operation [2], the operation proposed here is the cross-attention, developed for vision in [19]: this operation is directly focused on the extraction of associative information between both modality, from the IR features to the visible ones. For each stage of the features extraction, a module named **Cross-Attention Features Fusion (CAFF)** performs a hierarchical cross-attention from IR features to visible features.The operations performed on the features maps, at each level of abstraction, are illustrated in Figure 3. As shown, cross-attention is performed with multiple kernel size, in order to insist on features extraction at multiple scales. Then the correlation information obtained is combined with mono-modal features.

For n features maps extracted by each backbone, the proposed implementation is formalized as follows. Firstly a cross-attention is performed, correlating information from thermal modality in visible spectrum as described in Eq. 1. The operation is applied on features F_i^{thermal} and F_i^{visible}. The queries, keys and values matrices are defined using several convolution layers with different kernels sizes (hierarchical attention), increasing the capability of the model to extract multi-level correlations between the features from each modality.

$$\mathrm{CrossAttn}_i^k = \mathrm{softmax}\left(\frac{Q_i^k(F_i^{\text{visible}}).K_i^k(F_i^{\text{thermal}})^T}{\sqrt{d_k}}\right)V_i^k(F_i^{\text{thermal}}),$$
$$\forall i \in \{1,\ldots,n\}, \quad \forall k, \text{kernel size} \tag{1}$$

The cross-attention matrices obtained are stacked and compressed using convolution layer (depth C = 256). To enrich the extracted correlation features, several blocks of self-attention are applied on top of the fused cross-attention map. The correlation information is compressed again using a unitary convolution layer, in a depth of 64 channels and defined as $\mathrm{F}_i^{\text{Correlations}}$. Visible and thermal features maps are concatenated with $\mathrm{F}_i^{\text{Correlations}}$ to form $\mathrm{F}_i^{\text{Stacked}}$, in a depth of 2*C+64 channels as described in Eq. 2.

$$\mathrm{F}_i^{\text{Stacked}} = F_i^{\text{Visible}} + F_i^{\text{Thermal}} + \mathrm{F}_i^{\text{Correlations}},$$
$$\forall i \in \{1,\ldots,n\} \tag{2}$$

A 1*1 convolution is applied on the concatenated features that shapes the dimensionality of the vector to the shape attended by the transformer heads encoder in Eq. 3.

$$\mathrm{F}_i^{\text{Fused}} = \mathrm{Conv}_{2C+64\rightarrow C}^{k=1}(\mathrm{F}_i^{\text{Stacked}}),$$
$$\forall i \in \{1,\ldots,n\} \tag{3}$$

The proposed fusion module is optimized for the Swin-Large backbone and the DINO detector, and named CAFF (see Section 5.5 for head and backbones comparison). Table 1 describes the fusion module, the backbone type, the size of the convolution kernels used in the cross-attentions to compute keys, queries and values, and the number of additional attention blocks. The module CAFF corresponds to the module optimized with a Swin backbone, while CAFF* is optimized with a resnet50 and is compared with CAFF in the ablation study.

Fusion	Backbones	Cross-atn kernels	# Cross-atn block
CAFF	Swins	$k = \{1, 5, 7\}$	2
CAFF*	Resnets	k=1	3

Table 1. Description of the different fusion modules: CAFF and CAFF*.

5. Experiment

This section introduces the datasets used in the experiment, the training and evaluation setup, the performance obtained on aligned and misaligned images and the ablation study.

5.1. Data

Two public datasets defined as a set of IR-visible image pairs for object detection have been used to evaluate the performance of the architecture.

LLVIP dataset. LLVIP is a reference visible-infrared paired dataset dedicated to pedestrians detection for surveillance applications, mostly in low-light conditions [5]. The dataset contains 16.836 IR-visible image pairs, a majority of which were taken in low-light environments. All the couples of image of this dataset are strictly spatio-temporal aligned.

FLIR datasets. The FLIR-ADAS dataset is a multi-spectral object detection dataset including day and night scenes [20]. Three object categories are represented in the dataset: "person", "car" and "bicycle". **FLIR-aligned.** is the subset of this original dataset mainly used by the community, cleaned to contain only registered and paired images. The dataset contains 5,142 image pairs, which is a relatively limited amount of data considering transformers training data amounts and task difficulty.

5.2. Training and evaluation setup

Models initialization weights, for the backbones of each modality and the detection head, when available, come from COCO mono-spectral training. The fusion module is randomly initialized. Backbones are frozen during the training, for both modalities, preventing the over-parameterization of the whole model. Hyper-parameters such as the number of epoch, learning rate, loss are set following the original DINO recommendations [16]. The metric used for object detection performance evaluation is the mean average precision (mAP), calculated using Pycocotools package.

5.3. Object detection on aligned data

CAFF-DINO has been trained on both LLVIP and FLIR datasets. The comparison of the detection scores with state-state-of-the-art models is presented in Table 2 for LLVIP dataset, and Table 3 for FLIR-aligned dataset. These tables show the scores of each modality individually and fused. The combination Swin-Large-CAFF + DINO (called CAFF-DINO) reaches significantly higher performance compared with the state-of-the-art mono and multi-modality models from the literature. On LLVIP **an increase of mAP of 4.9 % is obtained** compared with CFT-YOLO-v5. On FLIR-aligned, **an increase of 9.1%** is obtained on the mAP compared with the ICA-Fusion model. The benefit of the information fusion between IR and visible is confirmed, considering the increase of performance between mono-modal DINO and multi-spectral proposal: it is measured an increase of mAP of 1 % on LLVIP (respectively 6.9 % on FLIR-aligned) between multi-modal DINO and IR only DINO. The lower benefit of CAFF-DINO on LLVIP

Dataset	Modality	Backbone	Detector	mAP50 (↑)	mAP75 (↑)	mAP (↑)
LLVIP	Visible	CSPD53	YOLOv5 [2]	90.8	51.9	50.0
	IR	CSPD53	YOLOv5 [2]	94.6	72.2	61.9
	Vis+IR	-	HalfWay [10]	91.4	60.1	55.1
	Vis+IR	-	ProbEn [21]	93.4	50.2	51.5
	Vis+IR	-	GAFF [11]	94.0	60.2	55.8
	Vis+IR	-	CSSA [8]	94.3	66.6	59.2
	Vis+IR	CFB	CFT-YOLO-v5 [2]	97.5	72.9	63.6
LLVIP **(our results)**	Visible	Swin-Large	DINO	91.3	59.8	54.4
	IR	Swin-Large	DINO	97.3	79.0	67.5
	Vis+IR	Swin-Large-CAFF	DINO	**98.1**	**79.0**	**68.5**

Table 2. Comparison between CAFF-DINO (swin-Large-CAFF+DINO) and several state-of-the-art fusion approaches on the LLVIP dataset.

Dataset	Modality	Backbone	Detector	mAP50 (↑)	mAP75 (↑)	mAP (↑)
FLIR-aligned	Visible	Resnet-50	Faster-RCNN	64.9	21.1	28.9
	IR	Resnet-50	Faster-RCNN	74.4	32.5	37.6
	Visible	CSPD53	YOLO-v5 [2]	67.8	25.9	31.8
	IR	CSPD53	YOLO-v5 [2]	73.9	35.7	39.5
	Vis+IR	ResNet18	GAFF [11]	72.9	32.9	37.5
	Vis+IR	-	ProbEn [21]	75.5	31.8	37.9
	Vis+IR	CFB	CFT-YOLO-v5 [2]	78.7	35.5	40.2
	Vis+IR	-	CSSA [8]	79.2	37.4	41.3
	Vis+IR	-	ICA-Fusion [22]	79.2	36.9	41.4
FLIR-aligned **(our results)**	Visible	Swin-Large	DINO	75.6	33.5	39.2
	IR	Swin-Large	DINO	77.2	41.3	43.6
	Vis+IR	Swin-Large-CAFF	DINO	**85.5**	**51.6**	**50.5**

Table 3. Comparison between CAFF-DINO (swin-Large-CAFF+DINO) and several state-of-the-art fusion approaches on the FLIR-aligned dataset.

could be explained by the properties of the data, with a majority of low-light acquisitions during the night. In this case, the IR information is the most beneficial and the fusion with visible spectra is less crucial.

Figures 4 and 5 give qualitative examples of detection on image pairs, from LLVIP and FLIR-aligned respectively, with ground-truth and network's detection (CAFF-DINO model). The model is able to detect efficiently objects in IR when they are more difficult to identify in visible spectrum (night situation). On the contrary, contrast between the different objects and the environment is more variable in IR in the FLIR pair (day-time, inducing more thermal saturation): here objects are easier to distinguish in visible. Confidence in the FLIR example is reduced due to the presence of numerous objects, which should explain the missed instances.

5.4. Object detection with misalignment

In order to evaluate the robustness of CAFF-DINO to error in image alignment (that can be caused by miscalibration or even no-calibration of the cameras), CAFF-DINO has been trained on the datasets with a systematic translation applied on the IR images. The misalignement has been applied on both train and test set. Performance of the proposed model is compared with CFT- YOLO-v5 model [2]. This model obtained the best performance in the state-of-the-art on LLVIP and is publicly available. A systematic translations of 10, 50, 100, and 200 pixels have been applied.

Tables 4 and 5 show the mAP and the relative mAP decrease associated with the systematic translation applied on LLVIP and FLIR-aligned. The mAP decrease is generally less important for CAFF-DINO, highlighting the robustness of the proposed architecture. The cross-attention operation could help the model to handle the misalignment between both modalities. The CFT-YOLO approach struggles more

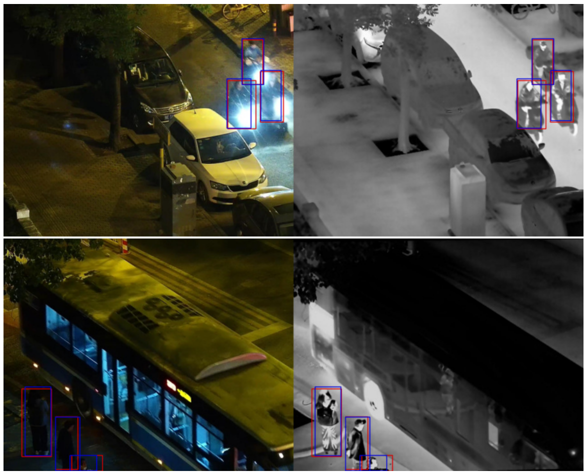

Figure 4. Visible and IR images from the LLVIP test-set. Ground-Truth are framed in red while CAFF-DINO detection are framed in blue (confidence threshold set to 50 %).

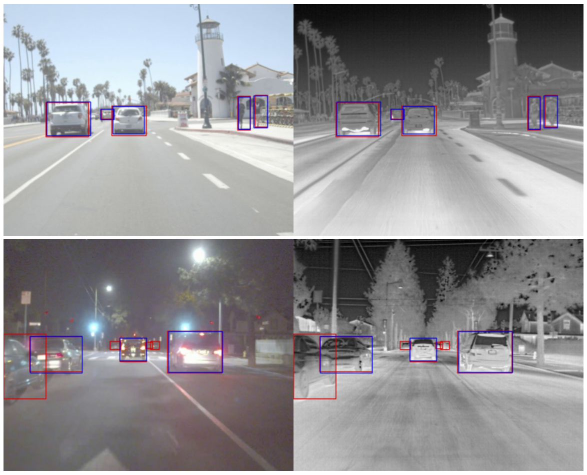

Figure 5. Visible and IR image pairs from the FLIR-aligned test-set. Ground-Truth are framed in red while CAFF-DINO detection are framed in blue (confidence threshold set to 50 %).

on misalignment in FLIR data which contains more challenging environment changes, requiring to perform accurate fusion of both spectra. As expected, both of the models studied here also converge to performance obtained in the unaltered spectrum (the visible spectrum) for the larger misalignment of 200 pixels, as more information is lost in IR, due to the systematic misalignment.

5.5. Ablation study

Several ablations are conducted in order to identify the contribution of each component of the architecture: transformer detector head, features extractors and features fusion approach.

Comparison of several transformers detectors. The performance of several heads are evaluated on LLVIP and FLIR in order to estimate the comparative benefit of our fusion approach with the different attention-based detectors elaborated in the literature. The detectors evaluated here are the original DETR [13], Lite-DETR [23], Deformable-DETR [14], H-Deformable-DETR [24] and DINO [16]. The experiment is conducted with Resnet-50 features extraction (CAFF*), due to the larger availability of public pre-trained weights with this backbone compared to Swin-Large. Table 6 shows the results of this experiment: the benefits of using modernized and improved transformer heads such as CAFF-DINO are confirmed. Detection performance on this dataset follows generally the technical progress of detection transformers, from the original DETR to DINO and modern architectures. These results tend to highlight the generic aspect of the proposed method to a large panel of detector head.

Comparison of different features extractors. The experiment is here conducted with Deformable-DETR, which provides public Resnet50, Swin-Tiny and Swin-Large pre-trained backbones associated to this architecture. Performance obtained with DINO are measured too, for comparison between CAFF and CAFF* (only Resnet50 and Swin-Large weights available).

Table 7 highlights the benefit of using richer features extractors such as transformers-based compared with Resnet. The performance gap between the Swin features extractor is more contrasted, both backbones giving comparable performance. In this paper, Swin-Large has been privileged in the CAFF-DINO model due its pre-trained weights availability. The scores using DINO also highlight the greater benefit of CAFF on Swin based backbone fusion (respectively CAFF* on Resnet based one).

Comparison of alternative features fusion approaches. Two alternative fusion modules are proposed as challengers of the CAFF fusion proposed. **Single features concatenation (concat)** fusion consists in concatenating the features maps from both spectra, at each level of abstraction. This fusion module is illustrated in Figure 6. The pipeline of this features fusion can be formalized as follows, for n features extractions performed by each mono-spectrum backbone:

The features maps from each spectrum, named respectively F_i^{Thermal} and F_i^{Visible} are stacked, giving a features vector of depth 2*C (Eq. 4). C is the features map depth attended by the transformer encoder.

$$\begin{aligned} \mathrm{F}_i^{\text{Stacked}} &= F_i^{\text{Visible}} + F_i^{\text{Thermal}}, \\ &\forall i \in \{1, \ldots, n\} \end{aligned} \tag{4}$$

A 1*1 convolution layer compresses this concatenated features vector to the shape attended by the transformer heads encoder in Eq. 5.

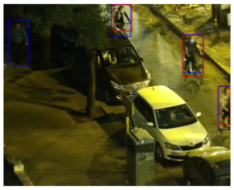				
Model	10px	50px	100px	200px
mAP(CFT-YOLO-v5)	51.3 (-19 %)	50.3 (-21 %)	49.2 (-23 %)	51.8 (-19 %)
mAP(CAFF-DINO)	57.4 (-16 %)	59.7 (-12 %)	57.0 (-17 %)	53.9 (-21 %)

Table 4. Illustration of the detection score (mAP) and the relative score decrease after misalignement on LLVIP. The relative score decrease is calculated as the relative difference between mAP on aligned data and misaligned data. The misalignment is indicated in pixel (px).

Model	10px	50px	100px	200px
mAP(CFT-YOLO-v5)	34.7 (-13 %)	28.7 (-28 %)	28.7 (-28 %)	29.0 (-27 %)
mAP(CAFF-DINO)	49.8 (-1 %)	37.2 (-26 %)	39.4 (-22 %)	43.7 (-13 %)

Table 5. Illustration of the detection score (mAP) and the relative score decrease after misalignement on FLIR-aligned. The relative score decrease is calculated as the relative difference between mAP on aligned data and misaligned data. The misalignment is indicated in pixel (px).

Dataset	Backbone	Head	mAP75	mAP
LLVIP	Resnet-50-CAFF*	DETR	65.5	58.9
		Deformable-DETR	69.4	61.1
		H-Deformable-DETR	75.5	65.0
		Lite-DINO	77.7	65.7
		DINO	**78.2**	**67.0**
FLIR-a	Resnet-50-CAFF*	DETR	20.6	15.7
		Deformable-DETR	36.4	**38.3**
		H-Deformable-DETR	33.7	34.6
		Lite-DINO	33.2	36.8
		DINO	**40.6**	37.9

Table 6. Comparison of different transformer heads from the literature combined with CAFF*1 on LLVIP and FLIR-aligned.

$$\mathrm{F}_i^{\text{Fused}} = \mathrm{Conv}_{2C\rightarrow C}^{k=1}(\mathrm{F}_i^{\text{Stacked}}), \quad \forall i \in \{1,\ldots,n\} \tag{5}$$

CAFF has also been compared with the **cosine-similarity operation (cos-sim)**, performed between the features maps from each spectrum [25, 26]. The cosine-similarity is a measure that calculates the similarity between the two features maps, by estimating the cosine of the angle between them. In Eq. 6, F_i^{Thermal} and F_i^{Visible} represent the mono-spectrum features maps. The operation can be viewed as a measure of the alignment between these IR and visible features, at each level of abstraction. A 1*1 convolution is added, reshaping the similarity matrix extracted for injection into the transformer's head.

Dataset	Head	Backbone	mAP75	mAP
LLVIP	Deformable-DETR	Resnet-50-CAFF*	69.4	61.1
		Swin-tiny-CAFF*	**76.2**	64.6
		Swin-Large-CAFF*	75.7	**64.9**
LLVIP	DINO	Resnet-50-CAFF*	76.3	66.3
		Swin-Large-CAFF*	78.1	67.6
		Resnet-50-CAFF	74.7	65.1
		Swin-Large-CAFF	**79.0**	**68.5**

Table 7. Comparison of several fusion backbones using CAFF* and CAFF combined with deformable-DETR and DINO on LLVIP.

$$\mathrm{F}_i^{\text{Correlations}} = \frac{F_i^{\text{Thermal}} \cdot F_i^{\text{Visible}}}{\|F_i^{\text{Thermal}}\| \times \|F_i^{\text{Visible}}\|}, \quad \forall i \in \{1,\ldots,n\} \tag{6}$$

Table 8 shows object detection scores obtained by the

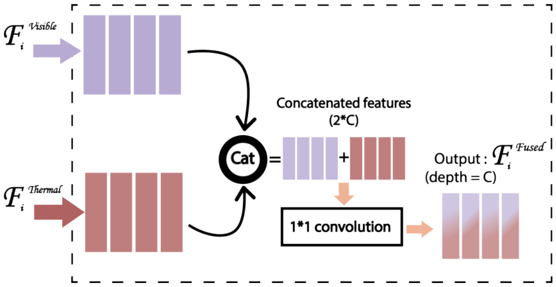
$\mathcal{F}_i^{Visible}$
$\mathcal{F}_i^{Thermal}$
Concatenated features
(2*C)
Cat
=
+
1*1 convolution
Output : $\mathcal{F}_i^{Fused}$
(depth = C)

Figure 6. Illustration of the fusion module approach with features concatenation only.

Dataset	Head	Backbone	mAP75	mAP
LLVIP	DINO	Resnet-50-CAFF*	**78.2**	**67.0**
		Resnet-50-concat	77.9	66.9
		Resnet-50-cos-sim	77.8	66.5
FLIR-aligned	DINO	Resnet-50-CAFF*	**40.6**	37.9
		Resnet-50-concat	35.1	**38.9**
		Resnet-50-cos-sim	38.3	36.6
FLIR-aligned	DINO	SwinL-CAFF	**51.6**	**50.5**
		SwinL-concat	48.6	48.4
		SwinL-cos-sim	42.6	43.7

Table 8. Comparison of different fusion approaches: CAFF/CAFF*, features concatenation and cosine-similarity on LLVIP and FLIR-aligned.

alternative modules on LLVIP and FLIR. It shows a decrease of performance by up to 6.8 % when using cosine-similarity on FLIR-aligned dataset (SwinL-cos-sim versus Swin-L-CAFF in the table) instead of the cross-attention proposal. If CAFF generally performs better, the gap between the different modules is reduced on LLVIP: the task is easier, with mostly night, low-light vision data favorable to IR modality. Single features concatenation (-concat in the table) gives high performance, close to CAFF and CAFF* (even outperforming the mAP with Resnet-50 features extraction on FLIR-aligned), highlighting the ability of transformers detectors to extract meaningful correlations directly on stacked then fused features maps from both modality. The single concatenation fusion might also give a lighter alternative module, considering the number of parameters. Figure 7 gives a qualitative illustration of the detection performance of the different fusion modules, on LLVIP: if the difference between CAFF (a) and concatenation (b) is limited considering box quality, there is a general decrease in detection confidence for the cosine-similarity fusion (c), inducing missed detection.

6. Conclusion

This work presents a method for multi-spectral object detection using a detection transformer combined with cross-attention features fusion. The proposed method is generic to the detection head. We show that the cross-attention module CAFF combined with DINO detection head outperforms the state-of-the-art models for object detection on several public IR-visible datasets, while being robust to the systematic misalignment applied on the IR images. In further works this architecture could be enriched with additional fusion operation, and spectrum-specific pre-training of the backbones instead of freezing. As the proposed method tends to be generic, it can be adapted to next generations of detection transformers. The extension of this work to more data starved IR-visible datasets, or to other visual modality fusion (visible-SAR, visible-LIDAR...) seems also relevant to estimate the generalization capabilities of the fusion method over the different spectra.

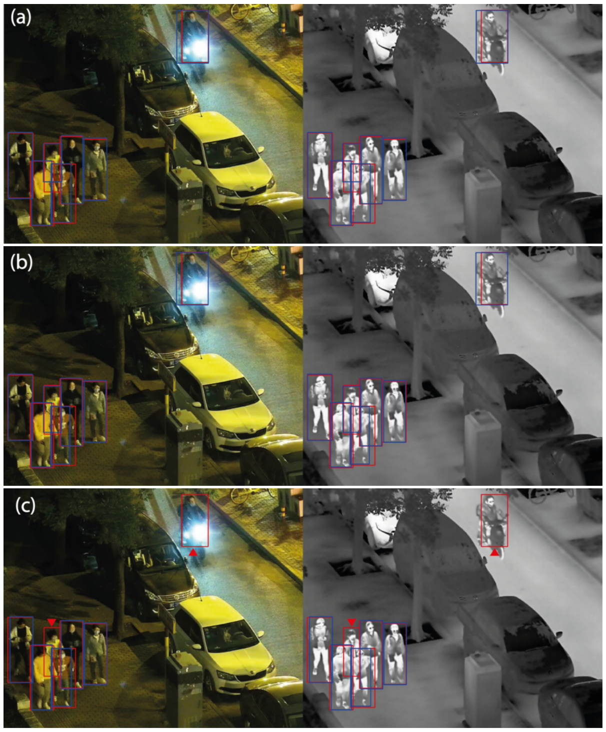

Figure 7. Visible and IR image pair from the LLVIP test-set. (a) is CAFF detection, (b) is concatenation, (c) is cosine-similarity. Ground-Truth are framed in red while our model's detection are framed in blue. Confidence threshold is 50 %. A red triangle indicates the missed detection.

Acknowledgments

The work reported in this study is supported in part by the French ministry of Armed Forces, *Agence de l'Innovation de Défense* (AID) for the PhD fellowship to Kevin Helvig.

References

[1] Soonmin Hwang, Jaesik Park, Namil Kim, Yukyung Choi, and In So Kweon. Multispectral pedestrian detection:

Benchmark dataset and baseline. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1037–1045, 2015. 1, 2

[2] Fang Qingyun, Han Dapeng, and Wang Zhaokui. Cross-modality fusion transformer for multispectral object detection, 2022. 1, 2, 3, 5

[3] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, and Feng Wu. Multi-modality cross attention network for image and sentence matching. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10938–10947, 2020. 1

[4] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models, 2022. 1

[5] 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. 1, 4

[6] Jörg Wagner, Volker Fischer, Michael Herman, and Sven Behnke. Multispectral pedestrian detection using deep fusion convolutional neural networks. In *The European Symposium on Artificial Neural Networks*, 2016. 2

[7] Jingjing Liu, Shaoting Zhang, Shu Wang, and Dimitris N. Metaxas. Multispectral deep neural networks for pedestrian detection. In Richard C. Wilson, Edwin R. Hancock, and William A. P. Smith, editors, *Proceedings of the British Machine Vision Conference 2016, BMVC 2016, York, UK, September 19-22, 2016*. BMVA Press, 2016. 2

[8] Yue Cao, Junchi Bin, Jozsef Hamari, Erik Blasch, and Zheng Liu. Multimodal object detection by channel switching and spatial attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 403–411, June 2023. 2, 5

[9] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85C:161–171, 2019. 2

[10] Heng Zhang, Elisa Fromont, Sébastien Lefevre, and Bruno Avignon. Multispectral fusion for object detection with cyclic fuse-and-refine blocks. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 276–280, 2020. 2, 5

[11] Heng Zhang, Elisa Fromont, Sebastien Lefevre, and Bruno Avignon. Guided attentive feature fusion for multispectral pedestrian detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 72–80, 2021. 2, 5

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 213–229, Berlin, Heidelberg, 2020. Springer-Verlag. 2, 6

[14] Zhu Xizhou, Su Weijie, Lu Lewei, Li Bin, Wang Xiaogang, and Dai Jifeng. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 2, 6

[15] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4794–4803, June 2022. 2

[16] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel Ni, and Heung-Yeung Shum. DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*, 2023. 2, 4, 6

[17] Liu Shilong, Li Feng, Zhang Hao, Yang Xiao, Qi Xianbiao, Su Hang, Zhu Jun, and Zhang Lei. DAB-DETR: Dynamic anchor boxes are better queries for DETR. In *International Conference on Learning Representations*, 2022. 2

[18] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6748–6758, October 2023. 2

[19] C. Chen, Q. Fan, and R. Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 347–356, Los Alamitos, CA, USA, oct 2021. IEEE Computer Society. 3

[20] FREE - FLIR Thermal Dataset for Algorithm Training | Teledyne FLIR, December 2023. [Online; accessed 17. Dec. 2023]. 4

[21] Yi-Ting Chen, Jinghao Shi, Zelin Ye, Christoph Mertz, Deva Ramanan, and Shu Kong. Multimodal object detection via probabilistic ensembling. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, page 139–158, Berlin, Heidelberg, 2022. Springer-Verlag. 5

[22] Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection, 2023. 5

[23] Feng Li, Ailing Zeng, Shilong Liu, Hao Zhang, Hongyang Li, Lei Zhang, and Lionel M. Ni. Lite detr: An interleaved multi-scale encoder for efficient detr. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18558–18567, June 2023. 6

[24] D. Jia, Y. Yuan, H. He, X. Wu, H. Yu, W. Lin, L. Sun, C. Zhang, and H. Hu. Detrs with hybrid matching. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19702–19712, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. 6

[25] Stijn van Dongen and Anton J. Enright. Metric distances derived from cosine similarity and pearson and spearman correlations, 2012. 7

[26] Takumi Nakagawa, Yutaro Sanada, Hiroki Waida, Yuhui Zhang, Yuichiro Wada, Kōsaku Takanashi, Tomonori Yamada, and Takafumi Kanamori. Denoising cosine similarity: A theory-driven approach for efficient representation learning, 2023. 7