YOLO-G3CF: Gaussian Contrastive Cross-Channel Fusion for Multimodal Object Detection

Abdelbadie Belmouhcine, Minh-Tan Pham, Sébastien Lefèvre

YOLO-G3CF: Gaussian Contrastive Cross-Channel Fusion for Multimodal Object Detection

Abdelbadie Belmouhcine, Minh-Tan Pham, and Sébastien Lefèvre

***Abstract*—Object detection is a crucial task in both computer vision and remote sensing. The performance of object detectors can vary across different modalities depending on lighting and weather conditions. To address these challenges, we propose a fusion module based on contrastive learning and Gaussian cross-channel attention, called Gaussian Contrastive Cross-Channel Fusion (G3CF). We integrate this module into a dual-YOLO architecture, forming YOLO-G3CF. The contrastive loss enforces similarity between the features sent to the detection head from both modality branches, as they should lead to the same detections. The Gaussian attention mechanism enables the model to fuse features in a higher-dimensional space, enhancing discriminative power. Extensive experiments on VEDAI, GeoImageNet, VTUAV-det, and FLIR demonstrate that G3CF improves detection performance, achieving a mAP increase of up to 6.64% over the best single-modality baselines and outperforming prior multimodal fusion methods. Regarding model complexity, our fusion method operates at a late stage, increasing the computational cost of single-modality YOLO by approximately 150% in terms of GFLOPs. For instance, YOLOv8 requires 52.84 GFLOPs, whereas YOLOv8-G3CF, due to its dual architecture and three G3CF modules, increases this to 131.22 GFLOPs. However, a single G3CF module requires only ~ 15 GFLOPs. Despite this overhead, our approach remains computationally less expensive than transformer-based models, e.g., ICAFusion requires 284.80 GFLOPs. Moreover, the proposed method still operates in real-time, achieving ~ 19 FPS on an NVIDIA RTX 2080. The code is available at https://github.com/abelmouhcine/YOLO-G3CF.**

***Index Terms*—Object detection, YOLO, multimodal fusion, Gaussian kernel, contrastive learning**

I. Introduction

Object detection continues to be a central focus in both computer vision [1] and remote sensing [2]. Thanks to advancements in deep learning [3], significant progress has been made in object detection for remote sensing, including aerial imagery. State-of-the-art detectors, whether single-stage or two-stage, anchor-based or anchor-free, and whether based on transformers or not, have been applied successfully to aerial images. However, they are predominantly designed for RGB images, making it difficult to exploit complementary information from other modalities such as thermal infrared (TIR), synthetic aperture radar (SAR), and LiDAR. Despite the demonstrated benefits of multimodal fusion, its application to object detection remains underexplored compared to segmentation [2].

For example, detectors based on RGB and TIR tend to outperform their single modality counter part [4]. Therefore, combining both visible and infrared images is crucial, with careful consideration of which modality should take precedence depending on the situation. When integrating both, it is essential to minimize redundancy between features from the two modalities, as duplication can degrade performance [5].

While existing multimodal object detectors have explored different fusion techniques, many rely on high-resolution images, focus primarily on small object detection, or introduce computationally expensive transformer-based architectures. To address these issues, we introduce YOLO-G3CF, a multimodal fusion method based on YOLO[1], designed to leverage Gaussian cross-channel attention while enforcing similarity between feature representations across modalities through contrastive loss. Indeed, contrastive learning is a representation learning approach that trains a model to distinguish between similar and dissimilar samples. It has proven to be effective and has garnered significant attention in various computer vision tasks [6], particularly in the context of self-supervised learning. Thus, our method employs contrastive loss to enforce feature alignment across modalities in a multimodal object detection framework. This ensures that both modality branches contribute consistently to the detection task. Besides, Ma et al. [7] used the Gaussian function in self-attention as a bias related to absolute positions to make the attention weights of HSI data distributions closer to the central query block. This proved useful for pixel classification.

Unlike previous work [8], [9], [10], our goal is not small object detection but rather a generalized and efficient fusion approach for multimodal detection. Moreover, we use the Gaussian function to compute cross-attention rather than position-based biases, since in our case, the images are already spatially aligned, and each pixel corresponds directly to the same position across modalities. Our key contributions are:

1) Contrastive loss to align features of the same object across modalities.
2) Cross-attention between channels instead of pixels, given spatial alignment between modalities.
3) Gaussian kernel for dot product in attention, enabling fusion in an infinite-dimensional Hilbert space.

Our experimental results demonstrate that the proposed fusion method outperforms several state-of-the-art multimodal object detection techniques on three remote sensing datasets (VEDAI [11], VTUAV-det [12] and GeoImageNet [13]) as well as one computer vision dataset (FLIR [14]), confirming the effectiveness of our approach. These improvements highlight the capability of our Gaussian Contrastive Cross-Channel Fusion (G3CF) module to leverage complementary information from different modalities.

All authors are with IRISA, Université Bretagne Sud, Vannes, France; Sébastien Lefèvre is also with UiT – The Arctic University of Norway, Tromsø, Norway.

Contact person: abdelbadie.belmouhcine@univ-ubs.fr

[1]https://github.com/ultralytics/ultralytics

The letter is structured as follows: Section 2 reviews prior fusion methods, Section 3 details our proposed method, Section 4 contains experiments and results, and Section 5 concludes the letter.

II. Related work

SuperYOLO [8] employs an encoder-decoder super-resolution module for pixel-level fusion of RGB and NIR images, particularly excelling at detecting small objects within vast backgrounds by leveraging high-resolution features through fusion. Building on this, Phung et al. [9] introduced SuperYOLOv8, an enhanced version of SuperYOLO, incorporating layers from YOLOv8 and utilizing Soft-NMS [15] for post-processing, outperforming previous state-of-the-art multispectral object detectors. Nevertheless, high resolution images are not always available for training. YOLOrs [10] also focuses on multimodal remote sensing images, providing oriented bounding boxes by applying six YOLOv5 layers per modality and fusing them using either concatenation or cross-product of feature maps. Both algorithms have demonstrated the effectiveness of multimodal fusion for object detection in remote sensing, particularly on the VEDAI dataset [11].

Zhang et al. [4] proposed a knowledge distillation framework called Cross-Modality HR Distillation (CMHRD), which facilitates the transfer of cross-modal information to a single-modality detector, targeting small object detection. However, transferring cross-modal information to a single modality does not allow the network to fully leverage the discriminative power of both modalities. Moreover, in this work, we do not specifically target small objects.

The integration of attention mechanisms has improved the performance of many computer vision algorithms for remote sensing tasks [16]. Fang and Wang [5] proposed a Cross-Modality Attentive Feature Fusion (CMAFF) approach, combining it with YOLOv5 to develop YOLOFusion. This method incorporates a Differential Enhancive Module and a Common Selective Module to amplify modality-specific features while selectively integrating shared ones to avoid redundancy. Their approach was compared against GAFF [17] and CFR [14] on the FLIR Dataset [14]. However, we believe that the use of spatial attention when modalities are aligned is not necessary. Indeed, the goal of spatial attention is to learn long-range dependencies. As the modalities are already spatially aligned, we know that each pixel in one modality corresponds to the pixel at the same position in the other modality. Therefore, there is no need to use the computationally expensive spatial attention in this case. Bismilla et al. [18] introduced an early-stage multimodal cross-channel model utilizing Swin transformers to improve spatial awareness. Their method computes cross-channel attention by processing each channel independently, yielding competitive results when compared to mid- and late-stage fusion approaches in the context of aerial image object detection. Besides, ICAFusion [19] introduces an iterative feature fusion module, leveraging a dual-branch backbone network to extract features from paired RGB and thermal images. The fusion process is driven by a dual cross-attention fusion transformer, which consists of three key components. First, the Spatial Feature Shrinking (SFS) mechanism, where the authors experimented with both convolution and mix-pooling to condense spatial features. Second, the Iterative Cross-modal Feature Enhancement (ICFE) module, where instead of stacking different blocks, they applied the same block iteratively within the transformer to refine feature fusion. Finally, they incorporated bimodal feature fusion to blend information from both modalities effectively.

The major drawbacks of those methods are that they either: 1) rely on high-resolution images, which are often unavailable and also, as the resolution increases, the number of GFLOPs also increases; 2) focus on small objects by stopping at higher scales and not going down to lower scales; 3) utilize transformers that require many iterations, making them computationally expensive. For instance ICA-Fusion has around $\sim 517M$ parameters and requires ~ 284 GFLOPs. We believe that these elements are not necessarily required when employing multimodal fusion, and that channel-wise attention is sufficient to effectively leverage information from different modalities.

III. Proposed Method

We propose a late fusion method that, unlike SuperYOLO [8], [9], does not rely on super resolution, which is not always available. Our approach leverages attention across channels rather than spatial attention, as the modalities are already spatially aligned. However, in contrast to [18], we compute cross-attention across all channels simultaneously rather than performing cross-attention in pairs of channels, and we do not use transformers. Computing attention on pairs of channels requires selecting pairs, which necessitates further empirical study. Moreover, the three RGB channels refer to the same modality since they are captured by the same sensor, and we aim to perform fusion between modalities from different sensors. Furthermore, transformers are computationally expensive as they repeat the attention computation multiple times, which is not always necessary.

Additionally, our fusion method enforces feature similarity between the two modalities. Since the fusion occurs in the later layers, the features from both modalities should ultimately converge to represent the same detection. While the modalities are already spatially aligned, enforcing feature similarity ensures that features extracted from different modalities are aligned in their representation, thus reducing modality gaps in representation. To achieve this, as shown in Fig. 1, we first compute the cosine similarity (CS) between aligned features F_1 and F_2, where F_1 and F_2 are matrices of shape $N \times d$. Here, N is the number of features (2D dimensions of feature maps are linearized to form N pixels), and d is their size (number of channels), with each row of the matrix corresponding to a spatial position. Initially, the cosine similarity is computed between F_1 and F_2 to measure the similarity of positive examples. Next, two random permutations Π_1 and Π_2 are applied to the first dimension (N) of F_1 and F_2, respectively, to calculate the cosine similarity for negative examples. Note that the cosine similarities are computed along d. Thus, CS has two outputs. The first one, marked as (0), is the cosine similarity between positive examples; features in the same

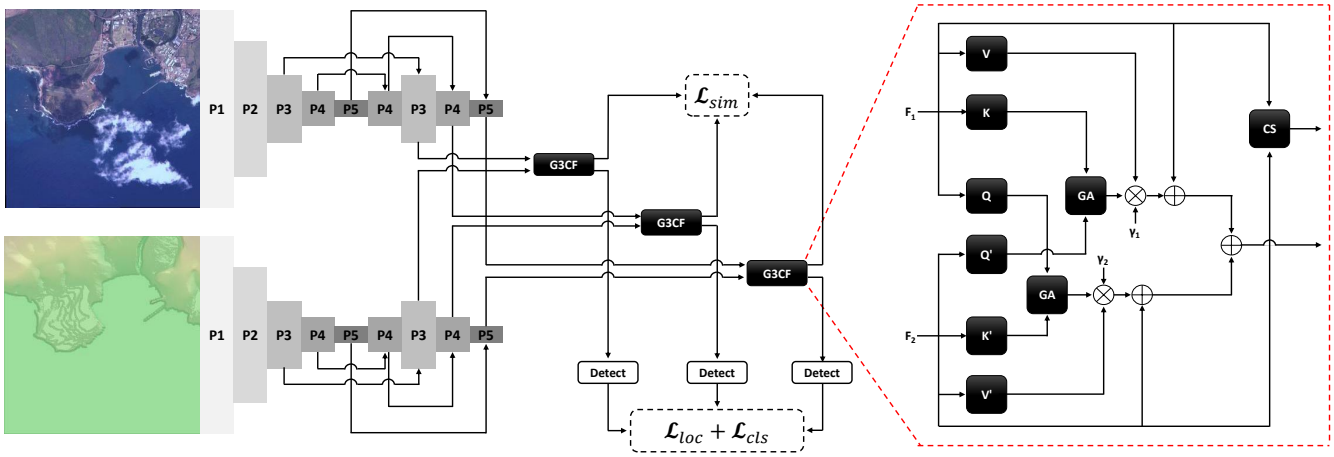


Fig. 1: Overall architecture of YOLO-G3CF. GA represents Gaussian Attention.

position in both feature maps. The second, tagged as (1), represents the cosine similarity between negative examples, which correspond to features whose positions have been permuted. After the permutation, we compute similarity only between features from different original positions that belong to the same class. The CS is defined as follows:

$$\mathrm{CS}(F_1, F_2) = \begin{cases} (0) & \max(\cos(F_1, F_2), 0) \\ (1) & \max(\cos(\Pi_1(F_1), \Pi_2(F_2)), 0) \end{cases} \tag{1}$$

To enable the attention mechanism to distinguish features specific to a given modality, we utilize a Gaussian kernel to compute attention in an infinite Hilbert space, rather than in the Euclidean space. The intuition behind this choice is to project the features into a higher-dimensional space, allowing the network to better capture and leverage the differences between modalities, while the contrastive loss ensures that the features from different modalities are aligned in their representation. We call our fusion method Gaussian Contrastive Cross-Channel Fusion (G3CF). The GA module in Fig. 1 represents Gaussian Attention between x and y, where (x, y) can be either $(K(F_1), Q'(F_2))$ for forward GA or $(K'(F_2), Q(F_1))$ for backward GA, and Q, K, V, Q', K' and V' are convolutional layers. GA corresponds to the following formula:

$$\mathrm{GA}(x, y) = e^{-\frac{A}{\gamma}}, \tag{2}$$

where γ is a hyper-parameter of the Gaussian kernel, A is the Gaussian attention map, $A_{ij} = ||x_i - y_j||_2^2$, and x_i, y_i are respectively the i$^{\text{th}}$ rows of matrices x and y. The fusion module is illustrated in Fig. 1 by the G3CF block. Two learned parameters, γ_1 and γ_2, control the contributions of forward and backward attentions between modalities during training.

We integrate our fusion module into YOLO by fusing the late layers of both modalities just before YOLO's head, resulting in the proposed model, YOLO-G3CF. The overall architecture of YOLO-G3CF is shown in Fig. 1. In addition to the standard YOLO losses ($\mathcal{L}_{loc}$ and $\mathcal{L}_{cls}$), we use a contrastive loss $\mathcal{L}_{sim}$, which is defined as follows:

$$\begin{cases} \mathcal{L}_{sim} = -\sum_N [(1 - y_0)^2 \times \log(y_0) + y_1^2 \times \log(1 - y_1)] \\ y_0 = \mathrm{CS}(F_1, F_2)(0) \\ y_1 = \mathrm{CS}(F_1, F_2)(1) \end{cases} \tag{3}$$

Similar to [9], we used Soft-NMS [15], which is well-suited for scenes with dense objects and significant overlap, a common scenario in remote sensing.

IV. Experiments

In this section, we present the datasets and evaluation metrics used to evaluate our proposed method, followed by detailed analysis and discussion.

A. Datasets

1) VEDAI: The VEDAI dataset [11] consists of 1,246 RGB/NIR image pairs. Only the 512 × 512 version is used in this study. The dataset includes 11 vehicle classes, but following [10], we excluded classes with fewer than 50 instances. We used 1,089 images for training and 121 for testing, following the 10-fold cross-validation protocol.

2) GeoImageNet: GeoImageNet [13] is a multi-source dataset for GeoAI applications, combining RGB aerial images and Digital Elevation Models (DEM). It comprises 876 image pairs across six natural feature categories We used the original train/test split with 698 training and 178 testing pairs. As a baseline, the authors applied Faster R-CNN and RetinaNet on GeoImageNet, concatenating mid-level features from two identical subnetworks before passing them through the rest of the network.

3) VTUAV-det: VTUAV-det [12], derived from VTUAV [20], is a drone-based person detection dataset featuring diverse scenes and weather conditions. This dataset contains pairs of RGB and TIR images. Following the original split, we used 11,392 training and 5,378 testing image pairs.

4) FLIR: The FLIR dataset [14] contains 5,142 aligned RGB/TIR image pairs, containing both daytime and nighttime scenes. In this letter, we used the aligned version [14], with 4,129 images for training and 1,013 for testing.

Fusion underlying technique		mAP@0.5 ↑	
Gaussian Attention	Contrastive Learning	YOLOv8	YOLOv11
✗	✗	85.9	88.1
✓	✗	84.8	89.3
✗	✓	87.4	89.1
✓	✓	**88.2**	**90.0**

TABLE I: Ablation study on GeoImageNet. In the absence of Gaussian Attention, Vanilla Attention is used.

Dataset	Modality	mAP@0.5 ↑	
		YOLOv8	YOLOv11
VEDAI	RGB	79.6	79.3
	NIR	73.1	74.2
	RGB+NIR (G3CF)	**81.0**	**80.4**
GeoImageNet	RGB	73.2	75.6
	DEM	86.1	84.4
	RGB+DEM (G3CF)	**88.2**	**90.0**
FLIR	RGB	68.8	69.9
	TIR	81.1	80.3
	RGB+TIR (G3CF)	**84.3**	**83.4**
VTUAV-det	RGB	40.1	39.6
	TIR	75.4	75.8
	RGB+TIR (G3CF)	**78.7**	**79.2**

TABLE II: Detection performance comparison of YOLOv8 and YOLOv11 on four datasets with single (without G3CF) and multi (with G3CF) modalities.

B. Evaluation metrics

We used AP@0.5 (Average Precision) and mAP@0.5 (mean Average Precision) to evaluate the performance of our module.

C. Ablation study

To evaluate the impact of our different contributions, we conducted an ablation study on the proposed G3CF module. As shown in Table I, the combination of Gaussian Attention with contrastive learning yields the best results on the GeoImageNet dataset. The use of contrastive loss to enforce similarity in feature representations across both modalities proves highly beneficial. Moreover, replacing the vanilla attention with Gaussian Attention enhances the model's ability to differentiate between modalities, as the attention is computed in a higher-dimensional space, allowing the network to capture more discriminative features.

To evaluate the effect of G3CF with respect to different versions of YOLO, we tested the fusion module with YOLOv8 and YOLOv11. As shown in Table II, the fusion has a similar effect on both versions, indicating that it can be generalized across different architectures. However, the more performant the baseline YOLO, the greater the performance of the fusion.

D. Results and discussion

We conducted experiments on four diverse datasets. In all cases, the proposed G3CF module outperforms other methods.

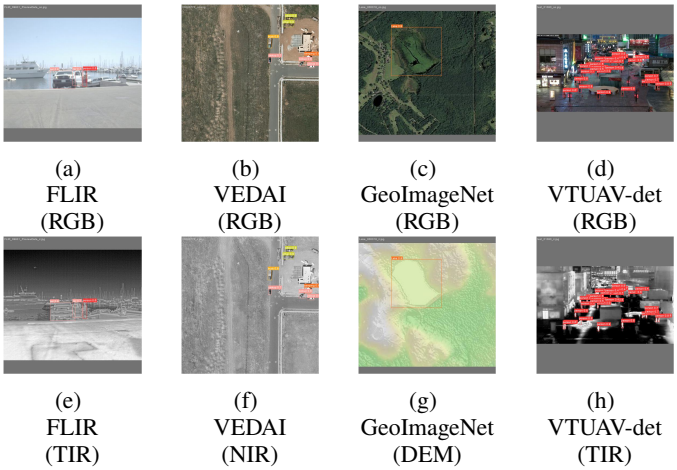
Table III presents the results on the VEDAI dataset. Our fusion method provides the best or second-best performance across categories. Additionally, as shown in Fig. 2, the NIR modality is not particularly useful in VEDAI, as all objects can be effectively detected from RGB images alone. Consequently, the performance of using only RGB is nearly equivalent to that of using RGB+NIR.

For GeoImageNet, as shown in Table IV, our fusion method effectively leverages both DEM and RGB, delivering either the best or second-best results across all categories. In this dataset, DEM proves beneficial for classes that do not involve water, which typically have low contrast, while RGB excels in classes with higher contrast. By combining both modalities, our method performs well across all classes. Furthermore, unlike SuperYOLO and YOLOrs, which are tailored for small objects, YOLO-G3CF is not hurt by large objects, maintaining strong performance regardless of object size.

As shown in Table V, our method achieves the best or second best results for both versions of YOLO and across all categories of the FLIR dataset. This is due to the nature of FLIR, which includes both day and night scenes and uses two modalities captured by different sensors. When these modalities are fused, the model is able to compensate for the false positives and negatives introduced by each modality individually. For instance, in the image shown in Fig. 2, using only the thermal modality results in detecting all objects with person-like shapes, which is not the case with RGB. Similarly, boats are sometimes misidentified as cars in RGB, but not in thermal images. By combining both modalities, these false positives are eliminated.

From Table VI, we observe the same behavior, with G3CF fusion outperforming QFDet [12], further demonstrating the effectiveness of the proposed approach.

As shown in Table III, our method nearly doubles the parameter count of single-modality YOLO due to its dual architecture and three G3CF modules. For instance, YOLOv8-G3CF reaches 131.22 GFLOPs, versus 52.84 GFLOPs for its single-modality counterpart. Yet, it remains lighter than Vision Transformer-based models like ICAFusion [19] (284.80 GFLOPs) and still run in real-time, achieving ~28 FPS for YOLOv8-G3CF and ~19 FPS for YOLOv11-G3CF on an NVIDIA RTX 2080. Note that for each method, GFLOPs are computed based on the input size used by the detector in the experiments, as specified in the second column of Table III. In case of resizing, the original aspect ratio is preserved.



(a) FLIR (RGB) (b) VEDAI (RGB) (c) GeoImageNet (RGB) (d) VTUAV-det (RGB)

(e) FLIR (TIR) (f) VEDAI (NIR) (g) GeoImageNet (DEM) (h) VTUAV-det (TIR)

Fig. 2: Prediction results of our method on four different datasets. Zoom in for more details.

Method	Size	Params ↓	GFLOPs ↓	AP@0.5 ↑								mAP@0.5 ↑
				Car	Pickup	Camping	Truck	Other	Tractor	Boat	Van	
SuperYOLO [8]	512	**4.85M**	**17.98**	91.1	85.7	79.3	70.2	57.3	80.4	60.2	76.5	75.1
YOLOrs [10]	512	-	-	84.2	78.3	68.8	52.6	46.8	67.9	21.5	57.9	59.7
Multimodal Transformer [18]	512	22.01M	87.89	89.1	82.7	76.4	61.6	56.3	77.9	60.4	75.8	72.5
YOLOFusion [5]	1024	12.50M	113.50	91.7	85.9	78.9	78.1	54.7	71.9	71.1	75.2	75.9
SuperYOLO8 [9]	512	8.73M	20.13	**93.2**	86.0	81.1	75.0	61.9	**86.4**	72.3	77.7	79.2
ICAFusion† [19]	1024	517.20M	284.80	-	-	-	-	-	-	-	-	76.6
YOLOv8-G3CF	512	113.55M	131.22	92.1	**86.2**	**83.0**	**81.8**	62.9	81.5	**82.2**	**78.5**	**81.0**
YOLOv11-G3CF	512	81.08M	87.55	90.8	86.0	81.0	80.8	**65.1**	82.7	80.9	75.8	80.4

TABLE III: Comparison on VEDAI dataset.[2]

Method	AP@0.5 ↑						mAP@0.5 ↑
	Bay	Lake	Island	Basin	Ridge	Valley	
RetinaNet‡ [13]	87.0	92.0	**99.0**	59.0	70.0	66.0	78.8
Faster-RCNN‡ [13]	79.0	80.0	98.0	38.5	58.0	71.0	70.8
YOLOrs*† [10]	-	-	-	-	-	-	13.4
SuperYOLO*† [8]	-	-	-	-	-	-	32.3
YOLOv8-G3CF	**91.1**	98.8	97.3	57.5	87.6	**97.0**	88.2
YOLOv11-G3CF	86.0	**99.1**	98.3	**70.0**	**90.1**	96.8	**90.0**

TABLE IV: Comparison on GeoImageNet dataset.[2]

Method	AP@0.5 ↑			mAP@0.5 ↑
	Bicycle	Car	Person	
GAFF† [17]	-	–	–	72.9
CFR [14]	57.8	84.9	74.5	72.4
CFT† [21]	-	-	-	78.7
ICAFusion [19]	66.9	89.0	81.6	79.2
YOLOv8-G3CF	**74.4**	89.8	**88.6**	**84.3**
YOLOv11-G3CF	71.4	**90.7**	87.9	83.4

TABLE V: Comparison on the FLIR dataset.[2]

Method	mAP@0.5 ↑
QFDet [12]	75.5
YOLOv8-G3CF	78.7
YOLOv11-G3CF	**79.2**

TABLE VI: Comparison on VTUAV-det dataset.[2]

V. Conclusion

In this letter, we have proposed a fusion method for object detection that leverages Gaussian cross-channel attention combined with contrastive learning for multimodal fusion. Our experiments on various datasets (VEDAI, GeoImageNet, VTUAV-det, FLIR) demonstrate that integrating G3CF into YOLO enables the detection model to fuse features more effectively by enhancing similar features through contrastive learning while capturing correlations across modalities via Gaussian cross-channel attention. While enhancing performance, this adds complexity but is still more efficient than transformer-based approaches like ICAFusion [19].

Our current fusion method works effectively with already spatially aligned modalities. However, in some scenarios, the modalities are not aligned. Hence, in the future, we plan to develop a fusion method for unaligned modalities, allowing the neural network to internally align features for a more robust fusion.

[2]**Bold** represents the best result, and underlined represents the second-best result. † Results per class were not provided in the original paper. ‡ Results are approximate, as they were extracted from a chart [13]. * Results are taken from [2].

References

[1] Z. Zou *et al.*, "Object detection in 20 years: A survey," *Proceedings of the IEEE*, vol. 111, no. 3, pp. 257–276, 2023.

[2] A. Belmouhcine *et al.*, "Multimodal Object Detection in Remote Sensing," in *IEEE International Geoscience and Remote Sensing Symposium*, 2023, pp. 1245–1248.

[3] Z. Li *et al.*, "Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey," *Remote Sensing*, vol. 14, no. 10, p. 2385, 2022.

[4] Y. Zhang *et al.*, "Learning Cross-Modality High-Resolution Representation for Thermal Small Object Detection," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[5] F. Qingyun *et al.*, "Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery," *Pattern Recognition*, vol. 130, p. 108786, 2022.

[6] H. Hu *et al.*, "A comprehensive survey on contrastive learning," *Neurocomputing*, p. 128645, 2024.

[7] C. Ma *et al.*, "Light Self-Gaussian-Attention Vision Transformer for Hyperspectral Image Classification," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–12, 2023.

[8] J. Zhang *et al.*, "SuperYOLO: Super resolution assisted object detection in multimodal remote sensing imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–15, 2023.

[9] K.-P. Phung *et al.*, "SuperYOLO8: Enhancing Performance of Object Detection in Real-Time Multi-Modal Remote Sensing Imagery through SuperYOLO and YOLOv8," in *International Conference on Computing and Communication Technologies*, 2023, pp. 551–556.

[10] M. Sharma *et al.*, "YOLOrs: Object Detection in Multimodal Remote Sensing Imagery," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 1497–1508, 2021.

[11] S. Razakarivony *et al.*, "Vehicle detection in aerial imagery: A small target detection benchmark," *Journal of Visual Communication and Image Representation*, vol. 34, pp. 187–203, 2016.

[12] Y. Zhang *et al.*, "Drone-based RGBT tiny person detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 204, pp. 61–76, 2023.

[13] W. Li *et al.*, "GeoImageNet: a multi-source natural feature benchmark dataset for GeoAI and supervised machine learning," *GeoInformatica*, pp. 1–22, 2022.

[14] H. Zhang *et al.*, "Multispectral fusion for object detection with cyclic fuse-and-refine blocks," in *IEEE International conference on image processing*, 2020, pp. 276–280.

[15] N. Bodla *et al.*, "Soft-NMS–improving object detection with one line of code," in *IEEE International Conference on Computer Vision*, 2017, pp. 5561–5569.

[16] S. Ghaffarian *et al.*, "Effect of attention mechanism in deep learning-based remote sensing image processing: A systematic literature review," *Remote Sensing*, vol. 13, no. 15, p. 2965, 2021.

[17] H. Zhang *et al.*, "Guided attentive feature fusion for multispectral pedestrian detection," in *IEEE Winter Conference on Applications of Computer Vision*, 2021, pp. 72–80.

[18] B. Bahaduri *et al.*, "Multimodal Transformer Using Cross-Channel attention for Object Detection in Remote Sensing Images," in *IEEE International Conference on Image Processing*, 2024.

[19] J. Shen *et al.*, "ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection," *Pattern Recognition*, vol. 145, p. 109913, 2024.

[20] P. Zhang *et al.*, "Visible-thermal UAV tracking: A large-scale benchmark and new baseline," in *IEEE / CVF Computer Vision and Pattern Recognition Conference*, 2022, pp. 8886–8895.

[21] F. Qingyun *et al.*, "Cross-modality fusion transformer for multispectral object detection," *arXiv preprint arXiv:2111.00273*, 2021.