2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

SAFL-Geo: Structure-Aware Feature Learning with Fusion Loss for Infrared-Visible Geo-Localization

Jiabo Shen[1], Shuying Zhao[1*], Yunzhou Zhang[1], Tengda Zhang[1], Hongyu Zhou[1], Yu Zhang[1], Jiaxu Gao[1]

***Abstract*— Cross-modal Visual Geo-localization often aims to retrieve a satellite visible-light image of the same geographic location from a large-scale database using an infrared image captured by an unmanned aerial vehicle (UAV), thereby achieving precise localization. This capability is crucial for autonomous drone localization and navigation in low-light conditions such as nighttime or smoky environments. However, research in this field is still in its nascent stage, with existing methods being few in number and limited in precision. To address these issues, this paper proposes a structure-aware and fusion-loss constrained cross-modal geo-localization network (SAFL-Geo), which enhances the accuracy of cross-modal image retrieval. Specifically, we design a structure-aware module embedded into the network backbone, substantially enhancing the model's ability to perceive and extract cross-modally consistent structural features (such as road and building contours). Furthermore, we propose a feature enhancement and aggregation module that projects the refined multi-modal representations into a unified embedding space, effectively reducing the cross-modal representation gap while preserving discriminative semantic structures. Finally, we propose a fusion loss constraint strategy that constructs intermediate fused features as a "bridge" to constrain the distribution distances between infrared and fused features, as well as between visible and fused features, thereby indirectly mitigating the modality gap. Extensive experiments on the Boson datasets show that our SAFL-Geo achieves superior state-of-the-art performance.**

I. INTRODUCTION

Visual Geo-Localization refers to the task of determining geographic location by matching ground-level query images, often captured by an unmanned aerial vehicle (UAV), against a reference database of satellite imagery [1][2][3]. This capability is of significant practical value in applications such as military reconnaissance, disaster response, and automated delivery systems [4][5]. However, most existing methods [6][7][8] operate primarily within the visible spectrum, rendering them highly dependent on favorable lighting and weather conditions. Their performance degrades drastically in nighttime, haze, or strong-shadow environments, severely limiting their effectiveness for all-weather deployment. Meanwhile, thermal infrared imaging has attracted growing attention due to its unique advantages. Unlike optical imaging, it does not depend on visible light and thus maintains stable observation capabilities under complete darkness, smoke occlusion, or complex lighting conditions. This capability has motivated active research in thermal geo-localization via cross-modal retrieval between thermal infrared and visible light.

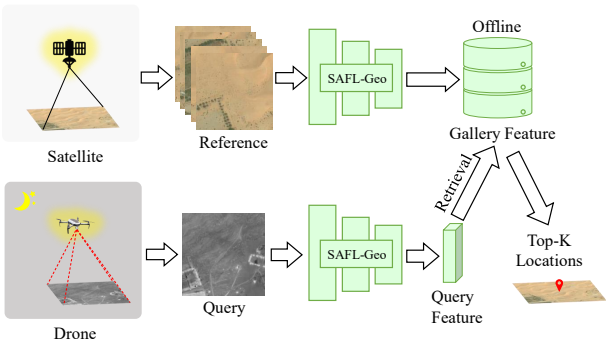

Fig. 1. **A cross-modal geo-localization method based on structural awareness and fusion loss constraints**. The reference database consists of RGB satellite images, while the query is an infrared image captured by a drone under low-light conditions.The retrieval process utilizes cosine similarity to compute the distance between the drone-based infrared features descriptors and the satellite image descriptors.

Thermal geo-localization plays a crucial role in enabling drone localization and navigation under low-visibility conditions such as nighttime or haze. Nevertheless, research in this area remains relatively scarce. A fundamental challenge hindering its development lies in the significant modality gap between thermal infrared images and the RGB satellite imagery that comprises most existing geo-localization databases. This discrepancy manifests not only in spectral properties but also in the loss or distortion of textural details. Prior research [9] attempted to directly adapt multiple single-modality methods to this cross-modal retrieval task, leading to suboptimal performance. A study [10] proposed a straightforward yet effective training-free aggregator for cross-modal retrieval tasks, but its accuracy remains low. These methods generally suffer from two major limitations. First, they tend to emphasize coarse-grained features like regional and background context, while overlooking fine-grained details such as object structures and contours—which are inherently stable across lighting variations, and thus highly valuable for cross-modal matching. Second, they commonly rely on minimizing feature distances in a shared embedding space, which becomes increasingly difficult to optimize due to the substantial modality divergence. Most existing cross-modal retrieval matching methods [11][12] attempt to bridge the gap through domain adaptation by generating intermediate or synthetic images for training. However, these techniques often introduce artifacts such as color inconsistency or reduced

*This work was not supported by any organization

[1]Jiabo Shen, Shuying Zhao, Yunzhou Zhang, Tengda Zhang, Hongyu Zhou, Yu Zhang, Jiaxu Gao are with College of Information Science and Engineering, Northeastern University, Shenyang 110819, China. zhaoshuying@ise.neu.edu.cn

This project is funded by National Natural Science Foundation of China(No. 61973066) and Major Science and Technology Projects of Liaoning Province (2021JH1/10400049).

image fidelity, further complicating the learning process and compromising localization accuracy.

To address these challenges, we propose SAFL-Geo, a novel cross-modal geo-localization network, as illustrated in Fig. 1. Our model is built upon a contrastive learning framework and lies in the synergistic use of three dedicated mechanisms—structural perception, fusion loss constraints, and feature aggregation—which collectively mitigate cross-modal discrepancies and enhance representation consistency. First, to leverage structurally stable information such as object contours that are often overlooked in existing methods, inspired by [13], we introduce a structure-aware constraint into the backbone feature extraction network. This enhances the model's ability to capture invariant features across thermal infrared and visible modalities. Second, we introduce a feature aggregation module that refines the features extracted by the backbone and projects them into a unified embedding space. This results in a more discriminative representation for downstream operations such as feature fusion and similarity measurement. Finally, to alleviate the difficulty of directly minimizing the feature distance between the two modalities, we propose a local fusion loss strategy. Central to this strategy is a cross-modal fusion module that serves as a "bridge" between modalities. This design avoids the hard alignment characteristic of traditional methods, thereby indirectly reducing the modality gap. Our main contributions are as follows:

- We propose a novel cross-modal geo-localization method called SAFL-Geo, which is based on a contrastive learning framework and introduces a structure-aware module to enhance the backbone network's extraction of cross-modal fine-grained stable features.
- We introduce a feature enhancement and aggregation module that strengthens modality-specific representations through multi-level transformations and projects them into a unified embedding space, thereby generating compact and highly discriminative features for downstream retrieval tasks.
- We design a local fusion loss strategy that adaptively reduces cross-modal discrepancies by leveraging fused features as a soft intermediary, thereby circumventing hard alignment between modalities.
- Both qualitative and quantitative experiments validate the effectiveness of our approach. Specifically, the proposed method achieves a R@1 of 80.55% on the public Boson dataset, surpassing previous state-of-the-art methods by over 8%.

II. RELATED WORK

A. Visual Geo-localization

In visual geo-localization tasks, effectively mitigating the perspective differences between UAV-captured and satellite images remains a fundamental challenge. Liu et al. [14] introduced directional geometric information into deep neural network training, thereby achieving better geo-localization results. Zhu et al. [15] used a Transformer-based approach with an attention-guided non-uniform cropping strategy to achieve excellent geo-localization results. Shao et al. [16] employed visual style transfer techniques to reduce cross-view differences. Deuser et al. [17] applied contrastive learning to cross-view geo-localization tasks. Chen et al. [18] built upon [17] by incorporating dynamic distance sampling strategies and spatial attention mechanisms, achieving the current state-of-the-art performance. Note that these methods rely on visible-spectrum cameras and exhibit significantly degraded performance under low-light conditions.

B. UAV Thermal Localization

Cross-modal geo-localization must address not only cross-perspective variations but also, more critically, the inherent discrepancies between different modalities. Early approaches, such as Thermal-Inertial Odometry [19], integrated thermal camera radiation measurements with inertial data to improve navigation accuracy in low-light conditions. To further bridge the modality gap, Gan et al. [20] introduced a multi-domain attention network that transfers knowledge from the RGB to the thermal infrared domain, facilitating visual navigation and localization at night and in adverse weather. Xiao et al. [21] utilized the Pix2Pix framework [22] with domain adaptation to exploit complementary information across modalities, proposing the first thermal infrared-to-visible cross-modal geo-localization method, which achieved state-of-the-art performance at the time.

In contrast to methods relying on domain adaptation, our approach formulates the problem within an image retrieval framework. We introduce a structure-aware module designed to extract modality-invariant structural features, and further employ joint constraints through feature fusion loss and enhanced aggregation mechanisms. This strategy effectively mitigates cross-modal discrepancies and significantly improves thermal infrared geo-localization accuracy.

III. METHOD

A. Overview

As shown in Fig. 2, the proposed SAFL-Geo framework comprises four core components: a backbone network, a structure awareness module, a feature enhancement aggregation module, and a feature fusion module. For each pair of visible and infrared images corresponding to the same geographic location, the inputs are first processed through the backbone network. We utilize a dual-stream architecture based on ConvNext [23] as the backbone. The structure awareness module enhances the model's capability to extract structural features, such as contours and edges, that remain consistent across modalities. These features are then passed to the feature enhancement aggregation module, which improves their discriminative power and projects them into a common feature space. Then, the feature fusion module integrates features from both modalities to generate a fused representation. To effectively bridge the modality gap, we introduce a fusion loss that constrains the distances among

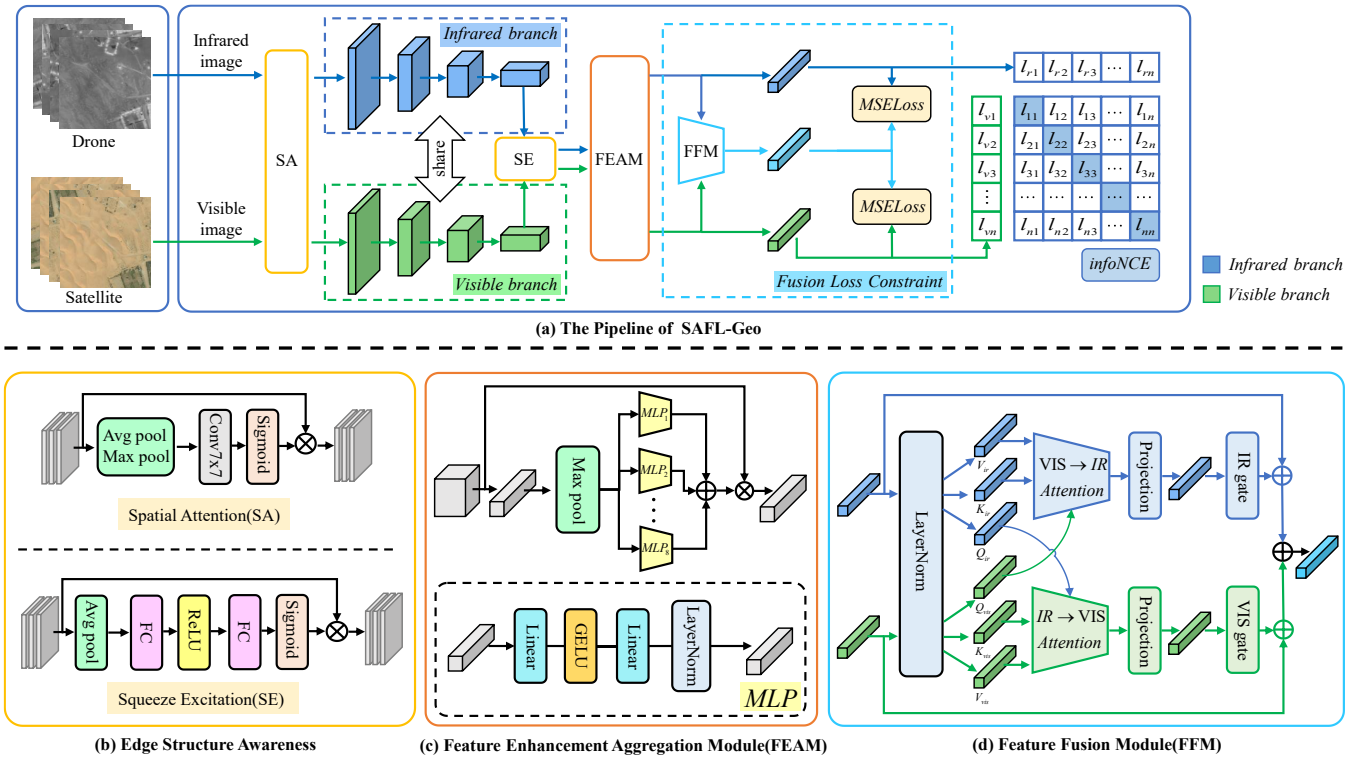

Fig. 2. **Overview of our proposed cross-modal thermal geolocation framework.** (a) The SAFL-Geo pipeline, a dual-stream network integrating structure-aware, feature aggregation enhancement, and feature fusion modules. (b) Edge structure-aware constraint comprises the Spatial Attention (SA) module and Squeeze-and-Excitation (SE) module. (c) The Feature Enhancement and Aggregation Module (FEAM) performs integrated feature refinement and projects representations into a unified embedding space. (d) Feature Fusion Module (FFM) generates fused features for subsequent fusion loss computation.

infrared features, visible features, and the fused representation through auxiliary loss terms, thereby progressively reducing cross-modal discrepancies.

B. Edge Structure Awareness

Images from different modalities exhibit significant differences in characteristics such as color distribution. However, structural information, particularly the contours of objects like buildings and roads, remains stable and consistent across modalities [24]. This makes structural cues highly reliable for cross-modal matching. To leverage this property, we designed a structure awareness module that enhances the network's ability to extract such modality-invariant structural features. We implemented this module using both spatial and channel attention mechanisms. The Spatial Attention (SA) module is integrated at the input stage of the backbone network to process each visible-infrared image pair prior to backbone feature extraction. As shown in Fig. 2(b), the SA module aggregates the average and maximum features across the channel dimension and applies a 7×7 convolution to generate a spatial attention map. This process directly enhances the contour and structural information of the input image by explicitly highlighting important regions during the preprocessing stage. The spatial attention mechanism is formally represented as follows:

$$M_s(X) = \sigma\left(f_{7\times7}\left(\left[AvgPool(X); MaxPool(X)\right]\right)\right) \quad (1)$$

where $X \in \mathbb{R}^{B\times C\times H\times W}$ is the input image, $AvgPool()$ and $MaxPool()$ represent average pooling and maximum pooling, respectively. $f_{7\times7}$ is a 7×7 convolution layer that performs convolution processing on the input image. σ is the sigmoid function, and $[\cdot;\cdot]$ represents channel concatenation. The SA module output is:

$$X_{SA} = X \odot M_s(X) \quad (2)$$

where $\odot$ denotes element-wise multiplication.

Subsequently, we incorporate a channel attention module at the output of the final stage of the backbone network. Specifically, we adopt the Squeeze-and-Excitation (SE) block [25], which comprises a squeeze step and an excitation step. In the squeeze step, global average pooling is applied to compress spatial information of each channel into a channel descriptor. Formally, this operation can be expressed as:

$$z_c = \frac{1}{H\times W}\sum_{h=1}^{H}\sum_{w=1}^{W} x_c(i,j) \quad (3)$$

where $x_c(i,j)$ is the value of the c-th channel at position (i,j), $z_c \in \mathbb{R}^{B\times C}$ represents the global information obtained for each channel, serving as the descriptor for that channel. The excitation phase captures inter-channel dependencies through a bottleneck structure formed by two fully connected layers, which adaptively recalibrates channel-wise feature responses. This process is formulated as:

$$s = \sigma\left(W_2\delta\left(W_1 z_c\right)\right) \quad (4)$$

where $s \in \mathbb{R}^{B\times C}$, $W_1 \in \mathbb{R}^{C/r\times C}$ and $W_2 \in \mathbb{R}^{C\times C/r}$ is the fully connected layer weight, δ is the ReLU activation

function, and r is the compression ratio. The output of the SE module is:

$$X_{SE} = s \cdot X_c \tag{5}$$

where $X_c \in \mathbb{R}^{B\times C\times H\times W}$ represents the final stage output features of the backbone network ConvNext.

The integration of these two components forms a dual attention mechanism operating across both channel and spatial dimensions, which significantly enhances the model's capacity to capture and represent edge structures present in different modalities.

C. Feature Enhancement Aggregation Module

To tackle the challenge of feature alignment in cross-modal infrared-to-visible image retrieval, we introduce a Feature Enhancement and Aggregation Module (FEAM), which is designed to project features from both modalities into a common latent space. As depicted in Fig. 2(c), FEAM first generates a spatial saliency map of the flattened features by max-pooling along the channel dimension. It then learns spatial-wise attention weights using parallel Multi-Layer Perceptron (MLP) branches, which maintain the independence of features across spatial dimensions. Each MLP branch utilizes a bottleneck structure composed of expansion and compression stages, and incorporates GELU activation along with normalization layers to facilitate stable gradient propagation and improve representational learning. The overall procedure can be formalized as follows:

$$Z_k = LayerNorm\left(W_2^k \cdot GELU\left(W_1^k \cdot X_{\max}\right)\right) \tag{6}$$

$$X_{sa} = stack\left(\left[Z_1, \cdots, Z_d\right], \dim = -1\right) \tag{7}$$

where $X_{\max} \in \mathbb{R}^{B\times(HW)}$ denotes the maximum value of the input feature along the channel dimension. W_1^k, W_2^k represent the weight parameters of the k-th MLP branch. $Z_k \in \mathbb{R}^{B\times C}$ is the output of the k-th MLP module. Stack denotes the operation of merging the outputs from multiple MLP branches. $X_{sa} \in \mathbb{R}^{B\times C\times d}$ represents the feature tensor obtained by aggregating all MLP outputs.

Subsequently, the original features are flattened into $X_{flat} \in \mathbb{R}^{\mathrm{B}\times\mathrm{C}\times(\mathrm{HW})}$ and multiplied with the spatial attention weights to enable interactive enhancement between spatial and channel information. Through transposition and reshaping operations, the module generates a compact joint feature representation, formulated as follows:

$$X_{out} = flatten\left(X_{flat} \cdot X_{sa}^T\right) \tag{8}$$

where $X_{out} \in \mathbb{R}^{B\times(d*C)}$ is the final output. This process effectively projects features from different modalities into a unified common space of fixed dimensionality while preserving critical spatial structural information.

D. Feature Fusion Module

To mitigate inter-modal discrepancies, we design a feature fusion module that acts as a "bridge" between infrared and visible features, thereby reducing the semantic gap between the two modalities. Inspired by cross-modal alignment perception [26], we employ a cross-attention mechanism [27] to facilitate feature fusion. Specifically, as shown in Fig. 2(d), we normalize the obtained infrared features $r \in \mathbb{R}^{B\times D}$ and visible features $v \in \mathbb{R}^{B\times D}$ through a LayerNorm layer to eliminate the scale differences between modalities, and obtain the processed features $\bar{r} \in \mathbb{R}^{B\times D}$ and $\bar{v} \in \mathbb{R}^{B\times D}$. We then apply a bidirectional cross-attention mechanism to model cross-modal feature interactions, formulated as follows:

$$A_{vis2ir} = softmax\left(\frac{Q_{vis} \cdot K_{ir}^T}{\sqrt{d_k}}\right) \cdot V_{ir} \tag{9}$$

$$A_{ir2vis} = softmax\left(\frac{Q_{ir} \cdot K_{vis}^T}{\sqrt{d_k}}\right) \cdot V_{vis} \tag{10}$$

where $Q_{ir}, K_{ir}, V_{ir} \in \mathbb{R}^{B\times H\times N\times d_k}$ are the query, key, and value vector of infrared features $\bar{r}$, $Q_{vis}, K_{vis}, V_{vis} \in \mathbb{R}^{B\times H\times N\times d_k}$ are the query, key, and value vector of visible features $\bar{v}$, and d_k is the dimension of each attention head.

The attention outputs from both directions are passed through a projection layer and adaptively weighted via a modality-specific gating mechanism using the Sigmoid function. This gate dynamically controls cross-modal fusion. The resulting features are then combined with the original input features through residual connections:

$$\hat{r} = r + g_r \odot Dropout\left(W_o A_{vis2ir}\right) \tag{11}$$

$$\hat{v} = v + g_v \odot Dropout\left(W_o A_{ir2vis}\right) \tag{12}$$

where g_r and g_v denote the gating modules for infrared and visible features, respectively. W_o is a learnable parameter matrix, and $\odot$ represents multiplication. Finally, we use average fusion to obtain the fused infrared-visible features.

E. Loss Function

To effectively bridge the gap between features of different modalities and mitigate cross-modal discrepancies, we introduce a fusion loss constraint strategy. As shown in Fig. 2(a), this approach augments the original baseline loss by incorporating an auxiliary fusion loss term that acts as a local constraint.

Our baseline loss function employs the symmetric InfoNCE loss [28][29], a widely adopted contrastive learning objective. In contrast to conventional contrastive losses, it enhances robustness by incorporating negative sample information from both modalities. The loss is defined as follows:

$$L_{infoNCE}\left(q, R\right) = -log\frac{\exp\left(q \cdot \frac{r_+}{\tau}\right)}{\sum_{i=0}^{R}\exp\left(q \cdot \frac{r_i}{\tau}\right)} \tag{13}$$

where q represents the query image, R represents a set of reference images, and r_+ is the positive sample uniquely corresponding to the query image q. During training, if the values between positive samples $q \cdot r_+$ are high and the values between negative samples are low, $L_{infoNCE}\left(q, R\right)$ approaches 0; otherwise, $L_{infoNCE}\left(q, R\right)$ grows exponentially. The hyperparameter τ is a learnable temperature parameter.

We use mean square error loss as the local fusion loss:

$$L_{MSE} = \frac{1}{2}\left(\|f_1 - f_{fused}\|_2^2 + \|f_2 - f_{fused}\|_2^2\right) \quad (14)$$

where f_1 and f_2 denote the infrared and visible light features, respectively. f_{fused} represents the fused feature, and $\|\cdot\|_2$ indicates the L2 norm.

We compute the distance errors between the infrared features and the fused features, as well as between the visible features and the fused features. Throughout the training process, these errors are iteratively minimized, thereby reducing the inter-modal feature distance. This process can be viewed as constructing a "bridge" between modalities, which effectively mitigates cross-modal discrepancies. To prevent feature collapse and ensure diversity among negative samples, we introduce an orthogonal regularization loss [30]:

$$L_{orth} = \lambda_{orth} \cdot \left\|Q \cdot K^T\right\|_F \quad (15)$$

where $Q \in \mathbb{R}^{m\times d}$ denotes the normalized query feature matrix, $K \in \mathbb{R}^{n\times d}$ denotes the normalized key feature matrix, $\|\cdot\|_F$ represents the Frobenius norm, and λ_{orth} is the regularization coefficient (set to 1×10^{-4} in our experiments). Minimizing $\left\|Q \cdot K^T\right\|_F$ will push the row vectors of Q and K toward orthogonality, thereby preventing feature collapse by promoting diversity across samples. The overall loss function is defined as:

$$L = \lambda_{global} \cdot L_{infoNCE} + \lambda_{local} \cdot L_{MSE} + L_{orth} \quad (16)$$

where λ_{global} and λ_{local} are hyperparameters to balance the loss during training.

IV. EXPERIMENT

A. Experimental Preparation

Datasets. Datasets containing cross-modal drone and satellite imagery for retrieval tasks remain scarce, with Boson [21] being, to our knowledge, the only publicly available collection of its kind. We therefore conduct all experiments using this dataset. This dataset comprises both raw and enhanced image data. The raw Boson thermal imaging dataset was captured from a near-vertical perspective with a spatial resolution of approximately one meter per pixel. It covers an area of 33 square kilometers, predominantly characterized by desert terrain, along with sparse farmland, roads, and buildings. The enhanced Boson dataset is generated through a domain adaptation method that synthesizes paired thermal infrared images from unaligned satellite visible images, thereby extending the original dataset.

Evaluation Metrics. We use Recall@K (R@K) and Average Precision (AP) to evaluate model performance. R@K measures the percentage of queries for which at least one correct database match (within a 25-meter radius) is found among the top-K retrieved results, while AP calculates the area under the Precision-Recall curve to consider both precision and recall.

Implementation Details. In our experiments, we adopt a ConvNeXt-Base backbone pretrained on ImageNet [31] as the base network, configured within a dual-branch architecture with weight sharing between the branches. Both infrared and visible input images are resized to 384×384 pixels. We use the Adam [32] optimizer with an initial learning rate of 1e-4, annealed following a cosine scheduling strategy. The model is trained for 80 epochs. During testing, cosine similarity is employed to compute the similarity between query and candidate images. All implementations are based on PyTorch [33], and experiments are conducted on an NVIDIA RTX 3090 GPU.

TABLE I

COMPARISON WITH STATE-OF-THE-ART METHODS ON BOSON DATASET. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Method	Drone → Satellite			Satellite → Drone		
	R@1	R@5	AP	R@1	R@5	AP
Sample4Geo[17]	21.13	29.13	21.97	25.02	36.72	24.91
ComplexUAV[18]	64.01	82.90	73.17	72.41	88.77	76.12
SGM ResNet-18[21]	72.33	84.96	75.65	73.41	86.67	76.34
Ours	80.55	96.56	83.13	82.80	97.96	82.01

TABLE II

COMPARISON WITH STATE-OF-THE-ART METHODS ON CONTRAST ENHANCEMENT BOSON DATASET. THE BEST AND SECOND-BEST RESULTS ARE HIGHLIGHTED IN RED AND BLUE, RESPECTIVELY.

Method	Drone → Satellite			Satellite → Drone		
	R@1	R@5	AP	R@1	R@5	AP
Sample4Geo[17]	64.35	70.13	56.92	66.15	75.88	63.84
ComplexUAV[18]	75.85	93.43	70.28	77.14	93.88	72.14
SGM ResNet-18[21]	92.10	96.90	83.88	93.33	96.94	83.97
Ours	97.14	99.45	86.22	98.68	99.94	87.91

B. Comparison with State-of-the-Art Methods

Results on the Boson Dataset. To evaluate the effectiveness of our approach, we conducted experiments under two retrieval modes: Drone → Satellite and Satellite → Drone. As shown in Table I, for these two retrieval tasks, our SAFL-Geo method surpasses SGM-ResNet-18 by 8.22% and 9.39% in R@1, and by 7.48% and 5.67% in AP, respectively. Given the limited number of methods available for cross-modal retrieval, we evaluated several single-modal approaches. Among these, ComplexUAV currently represents the state-of-the-art method for single-modal geo-localization. Nevertheless, it exhibits limitations when handling the significant disparities inherent in cross-modal imagery. Our approach effectively narrows this modal gap via an integrated strategy. In the drone-to-satellite retrieval task, it outperforms ComplexUAV by a notable margin, improving R@1 by 16.54% and AP by 9.96%. These results demonstrate that our method effectively mitigates the cross-modal discrepancy issue that challenges existing techniques, thereby enhancing the accuracy of thermal infrared geo-localization.

Results on the Contrast-enhanced Boson Dataset. As shown in Table II, although SGM-ResNet-18 achieves strong performance on the contrast-enhanced dataset, our method still surpasses it by 5.04% and 5.35% in R@1, and by

TABLE III

ABLATION OF METHODS EFFECTIVENESS ON BOSON DATASET. THE BEST PERFORMANCE ARE HIGHLIGHTED IN BOLD.

ESA	FEAM	Loss	Drone → Satellite			Satellite → Drone		
			R@1	R@5	AP	R@1	R@5	AP
×	×	×	65.44	84.17	74.72	73.14	89.74	76.44
✓	×	×	71.26	90.80	77.45	78.33	95.57	78.67
×	✓	×	73.16	91.14	79.73	76.57	92.13	77.20
×	×	✓	68.09	87.97	76.55	74.64	91.76	76.71
✓	✓	×	76.75	95.05	80.97	80.36	96.88	80.32
×	✓	✓	75.17	93.62	80.21	80.24	96.61	80.23
✓	×	✓	72.74	91.06	78.75	77.90	95.55	77.76
✓	✓	✓	**80.55**	**96.56**	**83.13**	**82.80**	**97.96**	**82.01**

TABLE IV

A COMPARISON OF DIFFERENT BACKBONE NETWORKS, INCLUDING THE NUMBER OF PARAMETERS, COMPUTATION, INFERENCE SPEED, AND R@K ACCURACY.

Backbone	Params	Macs	InferTime	R@1	R@5
ResNet50[34]	27.20M	24.50G	25.23ms	43.45	56.33
ConvNext-T[23]	30.24M	26.30G	25.99ms	77.96	94.22
ConvNext-S[23]	51.85M	51.16G	34.85ms	78.98	95.46
ConvNext-B[23]	91.84M	90.37G	42.28ms	80.55	96.56
ViT-S[35]	24.21M	25.00G	18.67ms	66.23	70.87
ViT-B[35]	88.01M	98.90G	21.78ms	70.04	79.33

2.34% and 3.94% in AP across the two retrieval modes. It also exhibits a more substantial advantage over single-modal methods. We attribute this improvement to the enhanced object contours and structural information in preprocessed data, which allows our structure-aware module to extract more discriminative features. These improved representations in turn facilitate more effective feature aggregation and fusion loss constraints, leading to further reduction of cross-modal discrepancies. The results confirm the superiority of our approach in thermal infrared geo-localization.

C. Ablation Studies

Effectiveness of Modules and Strategies. To evaluate the effectiveness of the proposed modules and strategies, we performed an ablation study on the Boson dataset. As summarized in Table III, the results demonstrate that employing either the Edge Structure Awareness (ESA) or the Feature Enhancement Aggregation Module (FEAM) module individually improves localization accuracy, while combining both leads to more substantial gains. This indicates that the ESA module enhances the model's ability to extract cross-modal stable features, while the FEAM module refines and projects features into a unified embedding space, thereby reducing inter-modal distance. As also observed in the table, introducing the local fusion loss brings additional improvements of 2.65% in R@1 and 1.83% in AP compared to the baseline. This confirms the effectiveness of the fusion loss during training. We argue that the fused features serve as a shared representation space, which acts as a transitional bridge between modalities and alleviates the challenges associated with direct feature alignment. When integrating all proposed components, our method achieves the best results with an R@1 of 80.55% and an AP of 83.13% in the Drone-to-Satellite retrieval mode, which represents the current best accuracy in thermal geo-localization. This fully demonstrates the rationality and effectiveness of our method.

TABLE V

THE IMPACT OF DIFFERENT INPUT SIZES ON MODEL PERFORMANCE.

Input	Drone → Satellite			
	R@1	R@5	R@10	AP
224×224	73.34	91.45	94.26	79.15
384×384	80.55	96.56	98.05	83.13
512×512	80.60	96.64	98.62	83.52

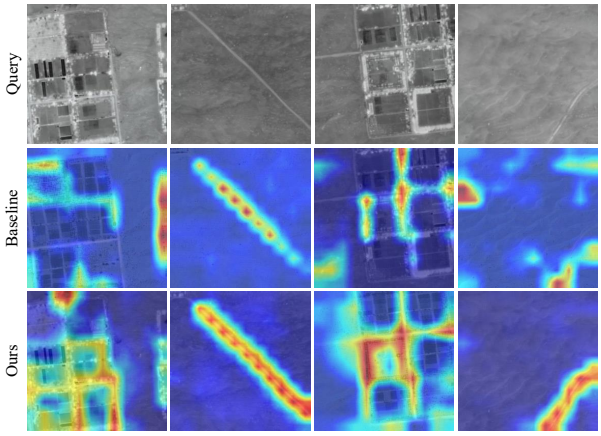

Fig. 3. **Visualization of feature heatmaps.** The top row shows the input infrared query image, while the second and third rows display the feature heatmaps produced by the baseline network and our proposed method, respectively. The results indicate that our approach more accurately concentrates on structurally stable components, such as building outlines and road contours.

Comparison of Different Backbone Networks. To explore the impact of different backbone networks on the UAV cross-modal thermal geo-localization task, we conducted experiments using several popular backbone networks, including ResNet50 [34], ConvNext-T, ConvNext-S, ConvNext-B [23], ViT-S, and ViT-B [35]. All models were trained on the Boson dataset with an input size of 384×384, and the tests were all Drone→Satellite retrieval results. The experimental results are shown in Table IV, including the number of parameters (Params), computational complexity (Mac), inference time (inference time for 2000 samples), and retrieval accuracy (R@1 and R@5). The experimental results show that the ViT series outperforms the ResNet series because ViT models have the ability to capture global dependencies in images. The ConvNext series of networks performs overall better, with ConvNext-B achieving the best results. We think the ConvNext network combines the advantages of ResNet and ViT, capturing global features while retaining the efficiency of convolution operations. In this paper, we use ConvNext-B as the base network.

Impact of Different Input Sizes. As shown in Table V, we evaluated the impact of input size on model performance using three common resolutions. The experiments were conducted on the Boson dataset. The experimental results indicate that the 384×384 input configuration performs better than its 224×224 counterpart. Although escalation to a 512×512 input size yields additional improvements, the gains

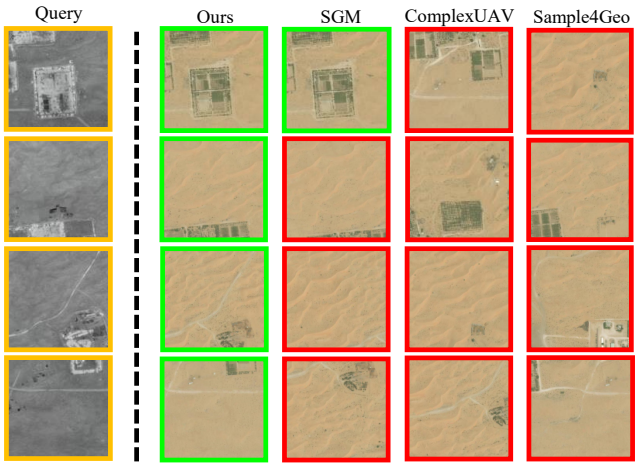

Fig. 4. **Visualization of retrieval results**. UAV-view query images are shown on the left of the dashed line, with the corresponding top-1 satellite retrieval results for each model displayed on the right. Correct matches are highlighted in green, and incorrect matches are marked in red.

are minimal and are achieved at the expense of significantly greater computational demands. Consequently, the 384×384 input size is utilized throughout all experiments in this paper.

D. Visualization

Visualization of feature heatmaps. To evaluate the feature representation capability of the proposed SAFL-Geo method, we visualized the corresponding feature maps. As shown in Fig. 3, the top row shows the input infrared query image, while the second and third rows display the feature heatmaps produced by the baseline network and our proposed method, respectively. Compared to the baseline method, our approach exhibits a stronger focus on structurally consistent elements such as roads and building contours. These results demonstrate the model's ability to extract modality-invariant features, confirming the effectiveness of our design.

Visualisation of retrieval results. In this section, we visualized the localization results of different models. Fig. 4 illustrates the retrieval results for four sample groups in the test set. The UAV view images (Query) are displayed to the left of the dashed line, while the R@1 results for each model in the satellite view are displayed to the right. True matches are indicated by green boxes, while false matches are indicated by red boxes. Through visualization, we observe that conventional single-modal retrieval algorithms struggle to effectively retrieve correct results, as modal differences limit their performance. The cross-modal retrieval algorithm SGM can correctly retrieve some query images but is significantly affected by similar images, resulting in low accuracy. In contrast, our method successfully retrieves all samples, demonstrating that our proposed strategies significantly mitigate modal differences and enhance positioning accuracy.

V. CONCLUSIONS

In this paper, we propose a novel cross-modal retrieval framework for thermal geo-localization, designed to address the significant challenges arising from modality discrepancies. By incorporating structure-aware constraints, our approach effectively captures stable and modality-invariant detailed features. We further introduce a feature enhancement and aggregation module that refines multi-modal representations and projects them into a unified embedding space, thereby preserving critical spatial and structural information. Moreover, a novel fusion loss strategy is proposed to bridge the modality gap by leveraging fused features as an intermediate representation, which substantially improves thermal geo-localization accuracy. Extensive experiments conducted on the Boson dataset and its enhanced variants validate that our method successfully mitigates cross-modal differences and achieves state-of-the-art performance. In future work, we will focus on further improving localization accuracy and exploring model lightweighting techniques to enable more efficient deployment.

REFERENCES

[1] Haoyang Wang, Fuhui Zhou, and Qihui Wu. Accurate vision-enabled uav location using feature-enhanced transformer-driven image matching. *IEEE Transactions on Instrumentation and Measurement*, 73:1–11, 2024.

[2] Panwang Xia, Yi Wan, Zhi Zheng, Yongjun Zhang, and Jiwei Deng. Enhancing cross-view geo-localization with domain alignment and scene consistency. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[3] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, 2021.

[4] Syed Agha Hassnain Mohsan, Nawaf Qasem Hamood Othman, Muhammad Asghar Khan, Hussain Amjad, and Justyna Żywiołek. A comprehensive review of micro uav charging techniques. *Micromachines*, 13(6):977, 2022.

[5] Syed Agha Hassnain Mohsan, Muhammad Asghar Khan, Fazal Noor, Insaf Ullah, and Mohammed H Alsharif. Towards the unmanned aerial vehicles (uavs): A comprehensive review. *Drones*, 6(6):147, 2022.

[6] Gaoshuang Huang, Yang Zhou, Luying Zhao, and Wenjian Gan. Cv-cities: Advancing cross-view geo-localization in global cities. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2024.

[7] Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, and Wenbo Hu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(9):4825–4839, 2023.

[8] Yuwen Yao, Cheng Sun, Tao Wang, Jianxing Yang, and Enhui Zheng. Uav geo-localization dataset and method based on cross-view matching. *Sensors*, 24(21):6905, 2024.

[9] Amulya Pendota and Sumohana S Channappayya. Are deep learning models pre-trained on rgb data good enough for rgb-thermal image retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4287–4296, 2024.

[10] Anuradha Uggi and Sumohana Channappayya. Training-free adapter for multi-modal image matching for all-day visual place recognition. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[11] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[12] Xiaoyan Qian, Miao Zhang, and Feng Zhang. Sparse gans for thermal infrared image generation from optical image. *IEEE Access*, 8:180124–180132, 2020.

[13] Yi Wang, Conrad M Albrecht, Nassim Ait Ali Braham, Chenying Liu, Zhitong Xiong, and Xiao Xiang Zhu. Decoupling common and unique representations for multimodal self-supervised learning. In *European Conference on Computer Vision*, pages 286–303. Springer, 2024.

[14] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019.

[15] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1162–1171, 2022.

[16] Jie Shao and LingHao Jiang. Style alignment-based dynamic observation method for uav-view geo-localization. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–14, 2023.

[17] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023.

[18] Junlong Chen, Gongjian Wen, Haojun Jian, and Xuxiang Fan. A visual localization benchmark for uavs in complex multi-terrain environments. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2025.

[19] Shehryar Khattak, Christos Papachristos, and Kostas Alexis. Keyframe-based direct thermal–inertial odometry. In *2019 International conference on robotics and automation (ICRA)*, pages 3563–3569. IEEE, 2019.

[20] Lu Gan, Connor Lee, and Soon-Jo Chung. Unsupervised rgb-to-thermal domain adaptation via multi-domain attention network. *arXiv preprint arXiv:2210.04367*, 2022.

[21] Jiuhong Xiao, Daniel Tortei, Eloy Roura, and Giuseppe Loianno. Long-range uav thermal geo-localization with satellite imagery. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5820–5827. IEEE, 2023.

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[23] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022.

[24] Cao Qin, Yunzhou Zhang, Yingda Liu, Delong Zhu, Sonya A Coleman, and Dermot Kerr. Structure-aware feature disentanglement with knowledge transfer for appearance-changing place recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 34(3):1278–1290, 2021.

[25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018.

[26] Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[27] Hezheng Lin, Xing Cheng, Xiangyu Wu, and Dong Shen. Cat: Cross attention in vision transformer. In *2022 IEEE international conference on multimedia and expo (ICME)*, pages 1–6. IEEE, 2022.

[28] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[29] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[30] Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12333–12343, 2021.

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[33] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[34] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.