AMCF-NET: ADAPTIVE MULTI-SCALE CROSS-MODAL FUSION NETWORK FOR UAV-SATELLITE CROSS-VIEW LOCALIZATION

Van Quan Ngo[1], *Quang Tung Pham*[1], *Chi Thanh Nguyen*[1,*]

Abstract

Cross-view localization between Unmanned Aerial Vehicle (UAV) and satellite imagery is crucial for autonomous navigation in GPS-denied environments. However, large domain gaps, including viewpoint discrepancies, scale variations, and appearance differences — pose significant challenges. In this paper, we propose the Adaptive Multi-scale Cross-modal Fusion Network (AMCF-Net), a novel approach that effectively addresses these limitations through a shared backbone architecture and adaptive fusion mechanisms. Unlike previous dual-backbone approaches that process UAV and satellite images separately, our method employs a unified FocalNet-Tiny backbone to extract cross-modal features, followed by a Spatially-adaptive Cross-modal Feature Fusion (AMCF) module that dynamically combines multi-scale similarities using learned adaptive weights. This shared representation learning enables better cross-modal alignment and significantly reduces computational overhead. Comprehensive experiments on the UL14 benchmark demonstrate that AMCF-Net achieves state-of-the-art performance, with a Relative Distance Score (RDS) of 78.12% and meter-level accuracy of 27.25% at 3 m, 50.16% at 5 m, 84.37% at 10 m, and finally 88.51% at 20 m. Ablation studies further validate the effectiveness of the shared backbone and adaptive fusion mechanism, demonstrating significant improvements over traditional separate processing approaches.

Index terms

UAV localization; satellite images; cross-view matching; multi-scale fusion; adaptive feature learning.

1. Introduction

Cross-view localization between UAV imagery and satellite maps underpins autonomous navigation in GPS-denied environments and supports applications in disaster monitoring and precision agriculture [1]. The core challenge lies in large domain gaps—severe viewpoint changes, scale disparities, and appearance variations between UAV and satellite views. This work focuses on point-to-point localization:

[1]Institute of Information Technology and Electronics, Academy of Military Science and Technology (AMST)
[*]Corresponding author, email: thanhnc@ioit.ai.vn
DOI: 10.56651/lqdtu.jst.v14.n02.1111.ict

given a UAV image captured from a specific GPS location at a fixed altitude (typically 80-100 m) with a nadir or near-nadir viewing angle, the task is to estimate the corresponding 2D pixel coordinates (x, y) of the UAV's capture location within a satellite reference image. This single-point correspondence objective reflects the practical scenario where each UAV image corresponds to one capture position, rather than establishing dense pixel-level correspondences across the entire image pair.

Early approaches cast the task as large scale retrieval by learning global descriptors for similarity search [2], [3]. While efficient, global pooling discards spatial details, limiting localization precision under strong viewpoint/scale changes [3]. Recent methods pursue dense, fine-grained matching to predict pixel-level correspondences [4], [5], but often incur high computational cost and rely on rigid, hand-crafted fusion, which weakens robustness to altitude- and content-induced scale mismatch [4], [6]. Hybrid pipelines attempt to balance coarse context and fine detail via hierarchical designs [6], [7], and shared backbones have been shown to reduce redundancy and improve alignment [8]. Despite this improvement, a key gap remains: adaptively fusing multi-scale, cross-modal evidence with spatial selectivity.

The paper address this gap with the AMCF-Net, which couples a *single* shared FocalNet-Tiny backbone with a lightweight, spatially-adaptive fusion module. The shared backbone enforces a unified representation for UAV and satellite inputs, improving cross-modal alignment while lowering compute. The fusion module computes multi-scale cross-view similarities and learns spatially varying weights to emphasize the most discriminative scale at each location, thereby reconciling local textures and global semantics under strong viewpoint changes.

The contributions are threefold: (1) a unified, parameter-shared backbone that cuts computation while improving cross-modal alignment; (2) a content-adaptive, spatially varying multi-scale fusion mechanism that selects scales per-pixel based on cross-view evidence; and (3) an empirical study on UL14 showing strong trade-offs between accuracy and efficiency (e.g., RDS 78.12%, MA@5 50.16%) together with ablations isolating the gains of sharing and adaptive fusion.

The remainder is organized as follows. Section 2 reviews related work. Section 3 details AMCF-Net. Section 4 presents experiments and ablations. Section 5 concludes and outlines future directions.

2. Related work

Existing UAV–satellite cross-view localization methods can be categorized into three approaches: large-scale matching, fine-grained matching, and hybrid methods with shared architectures.

Large-scale matching approaches rely on global descriptor learning through metric techniques. Early works by Vo and Hays [9] employed triplet loss [10] for ground-to-aerial matching, while Zheng *et al*. [2] introduced the University-1652 dataset with

Siamese networks [11]. Advanced pooling mechanisms, including NetVLAD [12], [13] and GeM pooling [14], improved global representations. Recently, Xu *et al*. [3] enhanced retrieval performance with context-aware descriptors and dynamic negative sampling. However, aggregation-based methods inherently sacrifice spatial information, limiting localization precision under significant viewpoint or scale variations.

Fine-grained matching approaches compute dense spatial correspondences to overcome global descriptor limitations. Dai *et al*. [4] pioneered pixel-level UAV–satellite matching with the DRL framework and UL14 dataset. Exploiting pyramidal architectures inspired by FPN [15], Wang *et al*. [6] developed WAMF-FPI with attention-weighted multi-scale fusion, while Fan *et al*. [5] introduced SSPT for iterative cross-view refinement through self-attention. Despite improved accuracy, these methods face high computational costs, rigid fusion strategies, and scalability challenges across diverse altitudes and perspectives.

Hybrid and shared backbone approaches combine complementary strengths while improving efficiency. Xu *et al*. [7] integrated coarse-to-fine pipelines with cross-attention, demonstrating unified processing advantages. He *et al*. [8] showed that DCD-FPI's single backbone with deformable convolutions reduces parameters by 50% while enhancing cross-modal alignment. However, existing methods employ fixed or hand-crafted fusion schemes that cannot adaptively weight multi-scale features based on spatial content. Focal modulation [16], which enables efficient multi-scale context modeling through hierarchical depth-wise convolutions, remains unexplored for cross-view matching with spatially-adaptive fusion.

The proposed AMCF-Net addresses this gap by leveraging FocalNet-Tiny [16] as a unified backbone enhanced with an AMCF module. Unlike prior fixed fusion strategies, AMCF dynamically learns spatially-varying weights, automatically emphasizing the most discriminative scale at each location based on content, thereby maximizing unified feature extraction benefits while enabling adaptive fusion.

3. The proposed method

3.1. Overall architecture

Given a UAV query image $\mathbf{I}^u$ and a satellite reference image $\mathbf{I}^s$, where H and W denote the spatial height and width of input images, the goal of the paper is to predict the UAV position in the satellite coordinate frame through cross-view correspondence. As illustrated in Fig. 1, AMCF-Net consists of three stages: shared feature extraction, adaptive multi-scale cross-modal fusion, and localization prediction. In Stage 1, the "Context Aggregation" block performs hierarchical contextualization via stacked depth-wise convolutions, and the "Modulator" block applies element-wise modulation to queries. The "gated aggregation" step (not explicitly labeled due to space) occurs between these blocks. In the AMCF module (Stage 2), the data flow is as follows:

(1) normalized features $\bar{\mathbf{F}}_i^u$ and $\bar{\mathbf{F}}_i^s$ are compared via cosine similarity to produce similarity maps $\mathbf{S}_i$, (2) $\mathbf{S}_0$ is fed into weight learning to generate adaptive fusion weights,

(3) the learned weights are applied to combine multi-scale satellite features $\bar{\mathbf{F}}_i^s$ through adaptive fusion, producing the final fused features $\mathbf{F}_{fused}^s$. Blue and red blocks denote UAV and satellite features, respectively.

Stage 1: Shared feature extraction. Both $\mathbf{I}^u$ and $\mathbf{I}^s$ are first embedded using a patch embedding module (STEM) and then fed into a single shared FocalNet-Tiny backbone. Sharing parameters allows both modalities to learn a unified representation space, promoting cross-modal alignment while reducing computation and model size. Unlike prior works that employ shared backbones for sensor fusion (e.g., camera-LiDAR from the same viewpoint), the approach addresses viewpoint-induced domain gaps where both inputs are RGB images but captured from drastically different viewing angles. The shared backbone forces both modalities through identical parameters, learning viewpoint-invariant geometric and semantic patterns that naturally reduce the domain gap between nadir UAV and overhead satellite views. Multi-scale features are extracted independently for each modality at three hierarchical stages: $\{\mathbf{F}_0^u, \mathbf{F}_0^s\}$ at resolution $H/4 \times W/4$, $\{\mathbf{F}_1^u, \mathbf{F}_1^s\}$ at $H/8 \times W/8$, and $\{\mathbf{F}_2^u, \mathbf{F}_2^s\}$ at $H/16 \times W/16$. No cross-modal fusion occurs in this stage.

Stage 2: Adaptive cross-modal fusion (AMCF). For each scale, normalized UAV and satellite features are aligned to the same spatial resolution and compared via cosine similarity, generating multi-scale similarity maps $\{\mathbf{S}_0, \mathbf{S}_1, \mathbf{S}_2\}$. A lightweight 1×1 convolution maps $\mathbf{S}_0$ to scale attention weights, followed by softmax to produce spatially-varying fusion weights. These weights adaptively combine multi-scale features, enabling the model to emphasize fine-grained textures or coarse semantic cues depending on local scene characteristics. The fused satellite feature map $\mathbf{F}_{fused}^s$ encodes the UAV location cues for final prediction.

Stage 3: Localization head. $\mathbf{F}_{fused}^s$ is projected through a lightweight channel embedding module to produce a coarse heatmap, which is upsampled to full resolution. The predicted UAV location (x_{pred}, y_{pred}) is the pixel with maximum response. This formulation provides a dense spatial likelihood map, improving robustness under ambiguous or repetitive patterns.

3.2. Shared feature extraction backbone

This section provides detailed implementation of Stage 1 (shared feature extraction). The backbone architecture employs hierarchical feature extraction, which is described as follows:

3.2.1. Focal modulation mechanism

The backbone employed in this study is FocalNet-Tiny [16], which replaces self-attention with focal modulation for efficiency and multi-scale modeling:

$$\mathbf{X}_{out}(i) = \mathbf{Q}(\mathbf{X}(i)) \odot \mathrm{FM}(\mathbf{X}_{ctx}(i)), \tag{1}$$

where, $\mathbf{X}(i)$ is the input token at spatial location i, $\mathbf{Q}$ is a linear query projection, FM denotes the focal modulation operation aggregating multi-scale context via learnable focal windows, and $\odot$ denotes element-wise multiplication (Hadamard product).

The focal modulation process (Fig. 1) consists of three steps:

(1) hierarchical contextualization (corresponding to the "Context Aggregation" block in Fig. 1) aggregates multi-scale context via stacked depth-wise convolutions with progressively larger receptive fields; (2) gated aggregation dynamically selects relevant scales using learned gating weights $\mathbf{g}$ (this process, though not explicitly labeled in the figure due to space constraints, occurs between hierarchical contextualization and element-wise modulation); and (3) element-wise modulation applies aggregated context through the Modulator $\mathbf{M}$ (corresponding to the "Modulator" block in Fig. 1) to queries via element-wise multiplication. The paper refers readers to [16] for detailed mechanisms, as these are internal to the FocalNet architecture.

3.2.2. Multi-scale feature extraction

To capture comprehensive multi-level semantic information, the paper extracts features from the first three hierarchical stages of the shared FocalNet-Tiny backbone. The paper employs FocalNet-Tiny-SRF [16], which consists of 4 stages with depths $[2, 2, 6, 2]$ and initial embedding dimension of 96. Focal modulation aggregates multi-scale context through hierarchical contextualization (stacked depth-wise convolutions), gated aggregation (learned scale selection), and element-wise modulation (context application to queries). The channel dimensions double at each stage: stage 0 has 96 channels, stage 1 has 192 channels, stage 2 has 384 channels, and stage 3 has 768 channels. Features are extracted from stages 0, 1, and 2 (output indices $[0, 1, 2]$):

$$\mathbf{F}_0^u, \mathbf{F}_0^s \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 96}, \quad \text{(low-level spatial detail)} \tag{2}$$

$$\mathbf{F}_1^u, \mathbf{F}_1^s \in \mathbb{R}^{\frac{H}{8} \times \frac{W}{8} \times 192}, \quad \text{(mid-level semantics)} \tag{3}$$

$$\mathbf{F}_2^u, \mathbf{F}_2^s \in \mathbb{R}^{\frac{H}{16} \times \frac{W}{16} \times 384}, \quad \text{(high-level semantic abstraction)} \tag{4}$$

where, B denotes the batch size and H, W represent the spatial dimensions of the input images. The shared FocalNet backbone processes both UAV and satellite images through the same parameter set, enabling unified representation learning that facilitates better cross-modal alignment. The paper initializes FocalNet-Tiny-SRF with ImageNet pretrained weights (focalnet_tiny_srf.pth). Although these weights are pretrained in single-modal image classification, the shared backbone learns cross-modal alignment through joint training on both modalities (detailed training strategy is described in Section 3.4).

3.3. AMCF Module: Adaptive Multi-scale Cross-modal Fusion

The AMCF module integrates multi-scale UAV and satellite features through similarity computation and adaptive fusion, establishing correspondences at multiple scales and

adaptively weighting their contributions. Given multi-scale features $\{\mathbf{F}_i^u, \mathbf{F}_i^s\}_{i=0}^2$ from the shared backbone, the AMCF process consists of three steps.

Step 1: Feature normalization. Channel dimensions are unified to 128 and spatial resolutions are aligned to $(H/4 \times W/4)$ via 1×1 convolution and bilinear interpolation, applied independently to both modalities:

$$\bar{\mathbf{F}}_i^u = \text{Resize}(\text{Conv}_{1\times1}(\mathbf{F}_i^u), (H/4, W/4)) \in \mathbb{R}^{B\times128\times H/4\times W/4} \tag{5}$$

$$\bar{\mathbf{F}}_i^s = \text{Resize}(\text{Conv}_{1\times1}(\mathbf{F}_i^s), (H/4, W/4)) \in \mathbb{R}^{B\times128\times H/4\times W/4} \tag{6}$$

This ensures all features operate in a common space (128 channels, $H/4\times W/4$ resolution) for meaningful cross-modal comparison.

Step 2: Multi-scale similarity computation. Normalized cosine similarity is calculated at each scale through element-wise multiplication of L2-normalized features:

$$\mathbf{S}_i = \frac{\bar{\mathbf{F}}_i^u}{||\bar{\mathbf{F}}_i^u||_2} \odot \frac{\bar{\mathbf{F}}_i^s}{||\bar{\mathbf{F}}_i^s||_2} \in \mathbb{R}^{B\times128\times H/4\times W/4} \tag{7}$$

where, $\odot$ denotes element-wise multiplication. The similarity maps ($\mathbf{S}_i$ in Fig. 1) are the outputs of this step, representing cross-modal correspondence signals at each scale. High similarity values at location (x, y) indicate potential UAV locations. Different scene structures match best at different scales: fine-scale features capture local texture details for precise matching, while coarse features capture high-level semantics robust to viewpoint changes.

Step 3: Adaptive fusion with learned weights. The model learns spatially-varying fusion weights from the finest similarity map $\mathbf{S}_0$, which contains the most detailed spatial information for precise scale assessment. $\mathbf{S}_0$ is chosen for weight learning because: (1) it has the highest spatial resolution ($H/4 \times W/4$), enabling precise location-specific scale selection; (2) it captures fine-scale texture patterns that are more viewpoint-invariant than high-level semantic features in coarser scales, providing stable alignment cues; and (3) its similarity values directly reflect the reliability of fine-scale matching, naturally indicating which scales should be emphasized at each location. The weight learning process first generates attention logits through a learnable 1×1 convolution:

$$\mathbf{A} = \text{Conv}_{1\times1}(\mathbf{S}_0; \boldsymbol{\theta}_{weight}) \in \mathbb{R}^{B\times3\times H/4\times W/4} \tag{8}$$

where, $\boldsymbol{\theta}_{weight} \in \mathbb{R}^{128\times3}$ is a learnable weight tensor (convolution kernel) of the 1×1 convolution layer that maps from 128 input channels (from $\mathbf{S}_0$) to 3 output channels (one per scale). These learnable parameters analyze similarity patterns to identify which scales provide reliable matching at each location. These logits are normalized via temperature-scaled softmax applied across the scale dimension:

$$\mathbf{W}(x, y, c) = [w^0(x, y), w^1(x, y), w^2(x, y)] = \frac{\exp(\lambda \cdot \mathbf{A}(x, y, c))}{\sum_{j=0}^{2} \exp(\lambda \cdot \mathbf{A}(x, y, j))} \tag{9}$$

where, $w^0(x, y)$, $w^1(x, y)$, and $w^2(x, y)$ are scalar fusion weights (elements of the weight tensor $\mathbf{W}$) at spatial location (x, y) for scales 0, 1, and 2, respectively.

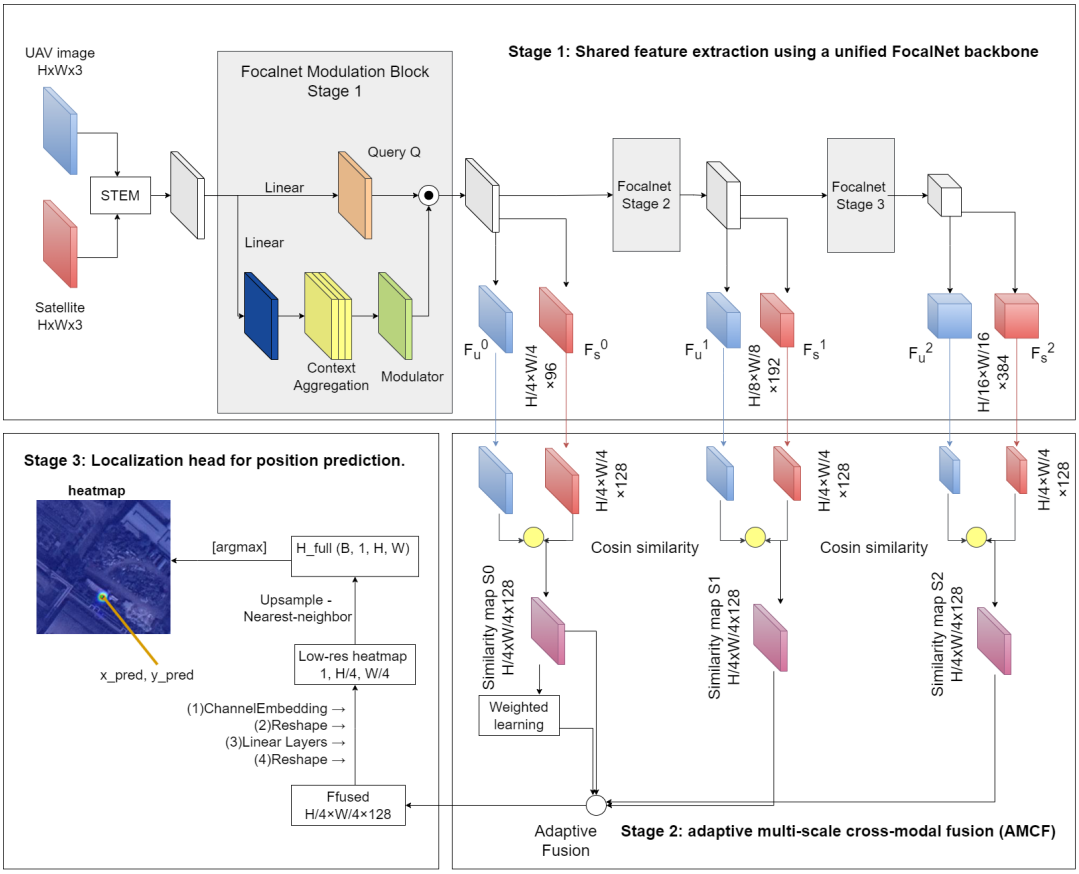

Fig. 1. Overall architecture of AMCF-Net. The framework consists of three stages: (1) shared feature extraction using FocalNet-Tiny backbone, (2) adaptive multi-scale cross-modal fusion (AMCF), and (3) localization head.

The temperature parameter $\lambda = 2.0$ controls attention sharpness (chosen through grid search over $\lambda \in \{0.5, 1.0, 2.0, 3.0\}$). The "adaptive" mechanism is realized through these spatially-varying weights: $\boldsymbol{\theta}_{weight}$ is learned during training, enabling the network to dynamically adjust scale emphasis based on local content characteristics at each spatial location (x, y), rather than using fixed weights across all locations. The learned weights are applied to combine normalized satellite features at all three scales:

$$\mathbf{F}^s_{fused}(x,y) = w^0(x,y) \cdot \bar{\mathbf{F}}^s_0(x,y) + w^1(x,y) \cdot \bar{\mathbf{F}}^s_1(x,y) + w^2(x,y) \cdot \bar{\mathbf{F}}^s_2(x,y) \quad (10)$$

The weights adaptively emphasize the most discriminative scale at each location: fine scales for distinctive textures, coarse scales for homogeneous regions. The fused features $\mathbf{F}^s_{fused}$ encode cross-modal correspondence through adaptive weighting, where high-similarity locations receive higher weights, emphasizing potential UAV locations.

3.4. Localization head and training loss

Localization head. The Channel Embedding head transforms fused satellite features $\mathbf{F}^s_{fused} \in \mathbb{R}^{B\times 128\times H/4\times W/4}$ into a heatmap $\mathbf{H}_{full} \in \mathbb{R}^{B\times 1\times H\times W}$. The transformation

consists of three operations: (1) reshape from spatial format $(B, 128, H/4, W/4)$ to token sequence $(B, H/4 \times W/4, 128)$ to enable per-location processing, (2) apply three linear projection layers that progressively reduce channel dimension $128 \to 64 \to 16 \to 1$ independently to each spatial token, and (3) reshape back to $(B, 1, H/4, W/4)$ and upsample to full resolution using nearest-neighbor interpolation. Each pixel value $\mathbf{H}_{full}(x, y) \in [0, 1]$ represents the confidence that the UAV location corresponds to pixel (x, y) in the satellite image coordinate system. During inference, the predicted location is obtained via argmax over the heatmap:

$$(x_{pred}, y_{pred}) = \arg\max_{(x,y)} \mathbf{H}_{full}(x, y) \tag{11}$$

where, (x_{pred}, y_{pred}) are pixel coordinates within the satellite image $\mathbf{I}^s$. Note that argmax is only used for inference to obtain discrete coordinates; during training, the loss function is computed directly on the continuous heatmap values $\mathbf{H}_{full}$ before argmax, enabling gradient-based optimization through backpropagation.

Training loss. The loss function is computed on the continuous heatmap $\mathbf{H}_{full}$ (before argmax), enabling gradient-based optimization through backpropagation. BalanceLoss addresses severe class imbalance: positive pixels (true UAV location) comprise $< 1\%$ of total pixels. The loss function is weighted binary cross-entropy with instance-wise normalization. For each training sample i, a binary target heatmap $\mathbf{Y}_i \in \{0, 1\}^{H/4 \times W/4}$ is constructed from ground truth location (x_{gt}, y_{gt}): the positive region $\mathcal{P}_i$ is a square with radius $R = 31$ pixels centered at (x_{gt}, y_{gt}) (all pixels labeled as 1), and the negative set $\mathcal{N}_i$ contains all other pixels (labeled as 0). Loss weights balance contributions:

$$w_i(p) = \begin{cases} \frac{1}{|\mathcal{P}_i|} & \text{if } p \in \mathcal{P}_i \\ \frac{\gamma}{|\mathcal{N}_i|} & \text{if } p \in \mathcal{N}_i \end{cases}, \quad \text{normalized: } \tilde{w}_i(p) = \frac{w_i(p)}{\sum_{p'} w_i(p')} \tag{12}$$

where, $\gamma = 130$ controls the relative importance of negative samples. The final loss is:

$$\mathcal{L}_{cls} = \sum_i \sum_p \tilde{w}_i(p) \cdot \text{BCE}(\mathbf{H}_i(p), \mathbf{Y}_i(p)) \tag{13}$$

where, $\mathbf{H}_i(p) = \sigma(\mathbf{H}_{full,i}(p))$ is the sigmoid activated prediction and $\text{BCE}(h, y) = -y \log(h) - (1 - y) \log(1 - h)$. The value $\gamma = 130$ was determined through empirical tuning to match the positive-to-negative ratio while maintaining training stability.

Training configuration. Optimization uses AdamW with initial learning rate 2×10^{-4} and cosine decay scheduling. To handle domain differences in the shared feature space and enable cross-modal alignment, we employ a mixed-batch training strategy where each mini-batch contains equal proportions of UAV and satellite images. This strategy ensures that gradient updates consider both modalities simultaneously, encouraging the shared backbone to learn domain-invariant representations that work effectively for cross-view matching. The mixed-batch approach is particularly crucial for the shared backbone architecture, as it prevents the network from overfitting to a single modality and facilitates unified representation learning that bridges the domain gap between UAV and satellite views.

Table 1. Performance comparison with state-of-the-art methods on UL14 dataset. Methods: FPI [4], WAMF-FPI [6], DRL [4], OS-FPI [7].

Method	MA@3↑	MA@5↑	MA@10↑	MA@20↑	RDS↑	FLOPs(G)↓
FPI [4]	-	18.30	38.36	57.67	57.22	14.88
WAMF-FPI [6]	12.49	26.99	52.63	69.73	65.33	13.32
DRL [4]	13.10	29.80	62.50	83.70	75.80	13.10
OS-FPI [7]	22.81	44.31	72.32	82.52	76.25	14.28
AMCF-Net (Dual-backbone)	24.78	47.69	78.91	86.11	77.32	16.52
AMCF-Net (Shared)	**27.25**	**50.16**	**84.37**	**88.51**	**78.12**	**12.30**

4. Experiments and results

4.1. Experimental setup

The paper evaluates the method on the UL14 benchmark dataset [4]. The UL14 dataset was originally constructed from 14 universities located in Hangzhou, China, containing 9,099 UAV images captured at three distinct flight altitudes (80 m, 90 m, and 100 m) with a consistent flight distance of 20 m between capture points. The original UAV images were center-cropped and resized to 512 × 512 pixels, while the corresponding satellite imagery was extracted from Google Earth based on GPS coordinates and initially processed at 1,280 × 1,280 resolution before being cropped to 384 × 384 pixels centered at the UAV's GPS location. The dataset exhibits several key characteristics: (1) dual-perspective imagery from both UAV and satellite viewpoints, (2) paired training data structure enabling supervised learning, (3) multi-altitude sampling providing diverse viewing angles and scale variations, and (4) multi-scale test configuration where test satellite images are constructed at 12 different spatial scales (ranging from 700 - 1,800 pixels at 100-pixel intervals, corresponding to approximately 0.294 meters per pixel), significantly increasing localization difficulty.

The training dataset consists of 6,768 paired UAV and satellite images gathered from 10 universities, averaging about 600 pairs per institution. The test set contains 2,331 UAV images from 4 additional universities with no geographical overlap with training locations, ensuring robust evaluation of cross-domain generalization. For each test UAV image, 12 satellite images at different scales are generated, resulting in a total of 27,972 satellite images in the test set. The UAV positions within satellite imagery are randomly distributed (center or edge positions), further increasing test complexity. The dataset presents significant challenges including viewpoint variations, seasonal changes, lighting conditions, and architectural diversity across different geographical locations.

This paper adopt standard metrics for cross-view localization evaluation. Meter-level Accuracy (MA@k) quantifies localization precision by measuring the percentage of test samples where the predicted location falls within k meters of the ground truth GPS coordinates. The spatial distance error (SD) is computed as:

$$\text{SD} = \sqrt{(\Delta x)^2 + (\Delta y)^2} \tag{14}$$

where, Δx and Δy denote the meter-level errors in longitude and latitude directions, respectively. MA@k is then defined as:

$$\text{MA@k} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\text{SD}_i<k} \tag{15}$$

where, N is the total number of test samples and $\mathbf{1}_{\text{SD}_i<k}$ is an indicator function that equals 1 when the spatial distance error for sample i is less than k meters, and 0 otherwise. Relative Distance Score (RDS) evaluates localization accuracy at the pixel level, providing scale-invariant assessment that is robust to different satellite image resolutions.

Unlike MA which depends on absolute spatial distances, RDS computes the relative pixel distance between predicted (X_p, Y_p) and ground-truth (X_g, Y_g) coordinates. The Relative Distance (RD) is first calculated as:

$$\text{RD} = \sqrt{\frac{(d_x/w)^2 + (d_y/h)^2}{2}} \tag{16}$$

where, $d_x = |X_p - X_g|$ and $d_y = |Y_p - Y_g|$ are pixel-level errors in horizontal and vertical directions, and w, h represent the width and height of the satellite image, respectively. RDS is then obtained through an exponential transformation:

$$\text{RDS} = e^{-k\times\text{RD}} = e^{-k\times\sqrt{\frac{(d_x/w)^2+(d_y/h)^2}{2}}} \tag{17}$$

where, $k = 10$ is a scaling factor that controls the sensitivity of the score to distance errors. RDS ranges from 0 to 1, with higher values indicating better localization accuracy. This metric offers three key advantages: (1) scale invariance through pixel-relative measurement, making it suitable for multi-scale test scenarios, (2) unified score representation that directly reflects model performance without requiring multiple threshold values, and (3) exponential decay that appropriately penalizes large distance errors, aligning with the expectation that significant deviations should be treated as localization failures.

The method was utilized using PyTorch and trained on a NVIDIA RTX 3060 GPU with 12 GB of memory. First, FocalNet-Tiny was initialized with ImageNet pretrained weights and fine-tuned end-to-end. Training used the AdamW optimizer with an initial learning rate of 2×10^{-4}, a cosine decay scheduler for 60 epochs, and gradient clipping with a norm of 1.0. The batch size was set to 8. Training convergence was monitored using loss and accuracy curves, which showed stable convergence after approximately 45 epochs.

4.2. Main results

Table 1 presents a comprehensive comparison between our AMCF-Net and state-of-the-art methods. The shared backbone approach achieves superior performance across all metrics, with MA@5 of 50.16%, MA@10 of 84.37%, and RDS of 78.12%.

Table 2. Ablation study showing progressive improvements of each component

Configuration	MA@5↑	MA@10↑	RDS↑	ΔMA@5
Dual-backbone baseline	47.69	78.91	77.32	–
+ Shared backbone	48.85	82.14	77.85	+1.16
+ Multi-scale AMCF	49.42	83.28	78.01	+0.57
+ Adaptive weights (Final)	**50.16**	**84.37**	**78.12**	**+0.74**

Notably, AMCF-Net (Shared) reduces computational complexity by 25.5% compared to our dual-backbone variant (12.30 vs. 16.52 GFLOPs) while achieving significantly higher accuracy through unified feature extraction. Compared to the best previous method OS-FPI, our approach improves MA@5 by 5.85 percentage points (50.16% vs. 44.31%) and MA@10 by 12.05 percentage points (84.37% vs. 72.32%).

Fig. 2 visualizes the comparison of MA@3, MA@5, MA@10, and MA@20 across all methods to visually highlight the performance improvement of AMCF-Net at different accuracy thresholds.

4.3. Ablation study

Ablation studies were conducted to validate key design choices. Table 2 shows the progressive improvements when adding each component. The key finding is that shared backbone outperforms dual-backbone by 2.47 percentage points MA@5 (50.16% vs 47.69%), demonstrating the effectiveness of unified feature learning. Additionally, the shared backbone approach achieves superior performance while reducing computational cost from 16.52 GFLOPs to 12.30 GFLOPs, representing a 25.5% reduction in computational complexity.

4.4. Efficiency analysis

Table 3 shows the AMCF-Net (shared) achieves optimal accuracy-efficiency balance (19 M parameters, 12.30 GFLOPs and 198 ms inference time). The shared backbone reduces computation by 25.5% while achieving highest RDS (78.12%).

Notably, the AMCF-Net (shared) achieves competitive inference time, being 1.23× faster than the baseline FPI method (198 ms vs 245 ms) and 1.2× faster than OS-FPI (198 ms vs 238 ms). This moderate speed improvement, combined with the highest accuracy (RDS = 78.12%), demonstrates a good balance between efficiency and performance for UAV localization applications. The shared backbone design achieves 25.5% FLOPs reduction while maintaining faster inference than dual-backbone approaches. For practical deployment on resource-constrained UAV platforms, AMCF-Net's 19 M parameters require approximately 76 MB memory (57% smaller than FPI's 178 MB), and the 198 ms inference time per image pair enables near-real-time localization suitable for autonomous navigation. The 12.30 GFLOPs computational cost is 17% lower than OS-FPI (14.28 GFLOPs), making AMCF-Net more suitable for embedded systems with limited resources.

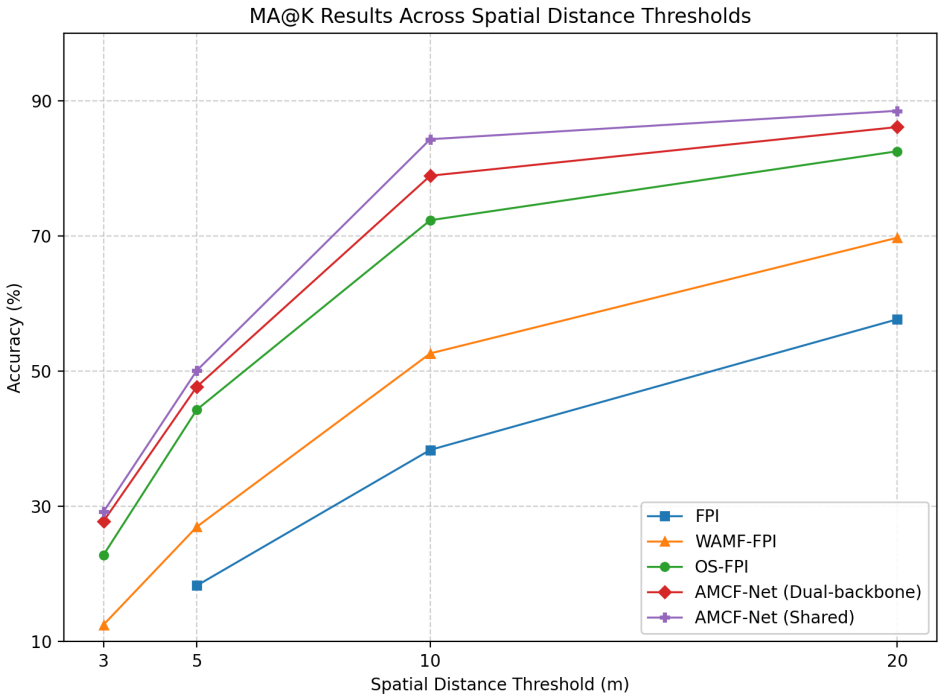

Fig. 2. Comparison of MA@3, MA@5, MA@10, and MA@20 across all methods.

Table 3. Efficiency comparison with different methods. Methods: FPI [4], OS-FPI [7], DRL [4]

Method	Params(M)↓	FLOPs(G)↓	Inference Time (ms) ↓	RDS↑
FPI [4]	44.48	14.88	245 (1×)	57.22
OS-FPI [7]	35.4	14.28	238 (0.97×)	76.25
DRL [4]	39.3	13.10	261 (1.07×)	75.80
AMCF-Net (Dual)	26.3	16.52	221 (0.90×)	77.32
AMCF-Net (Shared)	**19.0**	**12.30**	**198 (0.81×)**	**78.12**

4.5. Discussion

The shared backbone architecture with AMCF consistently improves performance across all metrics while significantly reducing computational overhead. The key insights from comprehensive analysis are: (1) shared backbone outperforms dual-backbone approaches by 2.47 percentage points MA@5 (50.16% vs 47.69%) while reducing computational cost by 25.5% (12.30 vs 16.52 GFLOPs), (2) learned spatially-varying weights within the unified representation space outperform average fusion, (3) multi-scale features extracted from the shared backbone are more discriminative than separate feature extraction, and (4) BalanceLoss with binary targets stabilizes training in the shared feature space.

The superiority of shared backbone can be attributed to three main factors: First,

unified feature learning enables better cross-modal alignment by forcing the network to learn common structural patterns. Second, parameter sharing reduces overfitting and improves generalization across different viewing conditions. Third, the reduced computational overhead allows for more complex fusion mechanisms without increasing the overall model complexity.

Comparison with SSPT [5] reveals interesting trade-offs: SSPT-384 achieves higher RDS (84.40% vs. 78.12%) through its single-stream pyramid transformer with self-attention and cross-attention mechanisms, but requires higher computational cost (15.28 vs. 12.30 GFLOPs) and more parameters (21.4 M vs. 19.0 M). SSPT-256 achieves 82.21% RDS with lower FLOPs (7.23 G), demonstrating the effectiveness of transformer-based architectures for cross-view refinement. It is worth noting that SSPT is designed for multimodal scenarios, incorporating style transfer technology to handle diverse environmental conditions including thermal infrared (TIR), nighttime, and rainy day datasets, whereas AMCF-Net focuses on standard RGB imagery under normal conditions. While AMCF-Net's RDS is lower, the proposed method offers superior parameter efficiency (19.0 M vs. 21.4 M) and maintains competitive FLOPs, making it more suitable for resource-constrained deployments. The key distinction lies in fusion strategies: SSPT employs fixed cross-attention for feature interaction, whereas AMCF-Net's adaptive spatially-varying fusion learns content-aware scale selection, potentially offering better generalization to diverse scene characteristics. The future work could explore combining SSPT's multimodal refinement mechanisms and environmental adaptation capabilities with AMCF-Net's adaptive fusion, potentially achieving both high accuracy, computational efficiency, and robustness across diverse imaging modalities and environmental conditions.

5. Conclusions

The paper has presented AMCF-Net, a novel approach for UAV-satellite cross-view localization that leverages a shared backbone architecture for superior performance and efficiency. The method addresses the critical limitations of existing dual-backbone approaches by employing unified feature extraction through a single FocalNet-Tiny backbone, combined with an adaptive multi-scale cross-modal fusion mechanism.

The comprehensive experimental analysis demonstrates several key findings. The unified architecture outperforms dual-backbone approaches by 2.47 percentage points in MA@5 (50.16% vs. 47.69%) while reducing computational overhead by 25.5% (12.30 vs. 16.52 GFLOPs), achieving optimal accuracy-efficiency trade-off. The AMCF module with learned spatially-varying weights consistently outperforms fixed fusion schemes, demonstrating the importance of content-adaptive scale selection for cross-view matching. AMCF-Net achieves the best results on the UL14 benchmark with 50.16% MA@5, 84.37% MA@10, and 78.12% RDS, outperforming the previous best method OS-FPI by 5.85 percentage points in MA@5. The shared backbone design reduces model parameters by 50% (from 39.6 M to 19.0 M) while achieving faster inference time (198 ms vs. 245 ms for baseline FPI).

The success of the approach can be attributed to three main factors: unified feature learning that enables better cross-modal alignment through shared representation spaces, parameter efficiency that allows deployment on resource-constrained platforms without sacrificing accuracy, and adaptive fusion mechanisms that dynamically adapt to varying scene contents and scale mismatches.

While superior performance is demonstrated on the UL14 benchmark, the evaluation is currently limited to a single dataset. Future work should validate generalization across additional datasets with diverse geographical characteristics (e.g., University-1652 [2], CVUSA) and varied viewing conditions. Moreover, extending the method to handle video sequences and integrating complementary sensors such as GPS and IMU could enhance robustness in real-world deployment scenarios. Further optimization for real-time applications, along with exploration of transformer-based fusion mechanisms, remains promising directions for future research.

References

[1] V. Q. Ngo, N. H. Long, P. H. Anh, T. T. T. Bui, and C. T. Nguyen, "Focal Hanning Loss: Revisiting the Heatmap Classification for UAV Self-Localization," in *Advances in Data Science and Optimization of Complex Systems*. Cham: Springer Nature Switzerland, 2025, pp. 452–461. DOI: 10.1007/978-3-031-90606-0_38

[2] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A Multi-view Multi-source Benchmark for Drone-based Geo-localization," in *Proceedings of the 28th ACM International Conference on Multimedia*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 1395–1403. DOI: 10.1145/3394171.3413896

[3] Y. Xu, M. Dai, W. Cai, and W. Yang, "Precise GPS-denied UAV self-positioning via context-enhanced cross-view geo-localization," *arXiv preprint arXiv:2502.11408*, 2025. DOI: 10.48550/arXiv.2502.11408

[4] M. Dai, J. Chen, Y. Lu, W. Hao, and E. Zheng, "Finding point with image: An end-to-end benchmark for vision-based UAV localization," *arXiv preprint arXiv:2208.06561*, 2022. DOI: 10.48550/arXiv.2208.06561

[5] J. Fan, E. Zheng, Y. He, and J. Yang, "A Cross-View geo-localization algorithm using UAV image and satellite image," *Sensors*, Vol. 24, No. 12, p. 3719, 2024. DOI: 10.3390/s24123719

[6] G. Wang, J. Chen, M. Dai, and E. Zheng, "WAMF-FPI: A weight-adaptive multi-feature fusion network for UAV localization," *Remote Sensing*, Vol. 15, No. 4, p. 910, 2023. DOI: 10.3390/rs15040910

[7] J. Chen, E. Zheng, M. Dai, Y. Chen, and Y. Lu, "OS-FPI: A Coarse-to-Fine one-stream network for UAV geolocalization," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, Vol. 17, pp. 7852–7866, 2024. DOI: 10.1109/JSTARS.2024.3380902

[8] Y. He, F. Chen, J. Chen, J. Fan, and E. Zheng, "DCD-FPI: A deformable convolution-based fusion network for unmanned aerial vehicle localization," *IEEE Access*, Vol. 12, pp. 129 308–129 318, 2024. DOI: 10.1109/ACCESS.2024.3415822

[9] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, 2016, pp. 494–509. DOI: 10.1007/978-3-319-46448-0_30

[10] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 815–823. DOI: 10.1109/CVPR.2015.7298682

[11] D. Chicco, "Siamese neural networks: An overview," *Artificial Neural Networks*, pp. 73–94, 2021. DOI: 10.1007/978-1-0716-0826-5_3

[12] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5297–5307. DOI: 10.1109/CVPR.2016.572

[13] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee, "CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7258–7267. DOI: 10.1109/CVPR.2018.00758

[14] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 7, pp. 1655–1668, 2019. DOI: 10.1109/TPAMI.2018.2846566

[15] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2117–2125. DOI: 10.1109/CVPR.2017.106

[16] J. Yang, C. Li, X. Dai, L. Yuan, and J. Gao, "Focal modulation networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 35, 2022, pp. 4203–4217. DOI: 10.48550/arXiv.2203.11926

Manuscript received 19-8-2025; Accepted 19-12-2025.

Van Quan Ngo is a researcher at the Institute of Information Technology and Electronics, AMST. He has contributed to many projects related to the application of information technology in promoting digital transformation and developing e-government systems in military units. His research interests focus on computer vision.
E-mail: ngoquanvnp@gmail.com

Quang Tung Pham is a researcher at the Institute of Information Technology and Electronics, AMST. He is currently pursuing his master's degree at Le Quy Don Technical University. His research interests include machine learning, computer vision and NLP, and the application of artificial intelligence to real-world problems.
E-mail: tungcntt55@gmail.com

Chi Thanh Nguyen is the Head of the AI Research Department at the Institute of Information Technology and Electronics, AMST. He received his Ph.D. degree in Computer Science from Nagaoka University of Technology, Japan, in 2012. His research interests include machine learning, computer vision and NLP.
E-mail: thanhnc@ioit.ai.vn

AMCF-NET: MẠNG HỢP NHẤT ĐA PHƯƠNG THỨC ĐA TỈ LỆ THÍCH ỨNG CHO BÀI TOÁN ĐỊNH VỊ CHÉO GIỮA ẢNH UAV VÀ ẢNH VỆ TINH

Ngô Văn Quân, Phạm Quang Tùng, Nguyễn Chí Thành

Tóm tắt

Bài toán xác định vị trí giữa ảnh chụp bởi phương tiện bay không người lái (UAV) và ảnh vệ tinh đóng vai trò quan trọng đối với việc điều hướng tự động trong các môi trường không có GPS. Tuy nhiên, có những thách thức lớn nảy sinh từ khoảng cách miền bao gồm sự khác biệt về góc nhìn, sự thay đổi về tỉ lệ và sự khác biệt về đặc điểm hình ảnh giữa ảnh UAV và ảnh vệ tinh. Trong bài báo này, chúng tôi đề xuất mạng hợp nhất đa tỉ lệ và đa phương thức thích ứng *(Adaptive Multi-scale Cross-modal Fusion Network – AMCF-Net)*, một phương pháp mới giải quyết hiệu quả các hạn chế này thông qua kiến trúc backbone dùng chung và cơ chế hợp nhất thích ứng.

Không giống như các phương pháp sử dụng hai backbone trước đây xử lý ảnh UAV và ảnh vệ tinh tách biệt, phương pháp của chúng tôi sử dụng một backbone FocalNet-Tiny thống nhất để trích xuất đặc trưng đa phương thức, tiếp theo là mô đun hợp nhất đặc trưng đa phương thức thích ứng không gian (AMCF) để kết hợp động các đặc trưng đa tỉ lệ dựa trên trọng số thích ứng được học. Cách tiếp cận học biểu diễn chung này giúp cải thiện đáng kể khả năng căn chỉnh đa phương thức và giảm đáng kể chi phí tính toán.

Các thí nghiệm toàn diện trên bộ dữ liệu UL14 cho thấy AMCF-Net đạt hiệu suất hàng đầu, với Điểm khoảng cách tương đối (RDS) đạt 78,12% và độ chính xác ở mức mét là 27,25% tại 3 m, 50,16% tại 5 m, 84,37% tại 10 m và 88,51% tại 20 m. Các thử nghiệm xác nhận hiệu quả của mạng backbone dùng chung và cơ chế hợp nhất thích ứng cho thấy những cải thiện đáng kể so với các phương pháp xử lý tách biệt truyền thống.

Từ khóa

Định vị UAV; ảnh vệ tinh; đối sánh đặc trưng đa góc nhìn; hợp nhất đặc trưng đa tỉ lệ; trọng số thích ứng.