A Vehicle-Mounted Radar-Vision System for Precisely Positioning Clustering UAVs

Guangyu Wu, *Member, IEEE*, Fuhui Zhou, *Senior Member, IEEE*, Kai Kit Wong, *Fellow, IEEE*, and Xiang-Yang Li, *Fellow, IEEE*

***Abstract*—The clustering unmanned aerial vehicles (UAVs) positioning is significant for preventing unauthorized clustering UAVs from causing physical and informational damages. However, current positioning systems suffer from limited sensing view and positioning range, which result in poor positioning performance. In order to tackle those issues, a novel vehicle-mounted radar-vision clustering UAVs positioning system is developed, which achieves precise, wide-area, and dynamic-view sensing and positioning of the clustering UAVs. Moreover, a matching-based spatiotemporal fusion framework is established to mitigate cross-modal and cross-view spatiotemporal misalignment by adaptively exploiting the cross-modal and cross-view feature correlations. Furthermore, we propose an attention-based spatiotemporal fusion method that achieves a trinity projective attention with the unique structure and task-oriented format for effective feature matching and precise clustering UAVs positioning. Our method also exploited the modality-oriented cross-modal feature and the UAV-motion-oriented cross-view UAV spatiotemporal motion feature.We demonstrate the advantages of our proposed framework and positioning method in our developed clustering UAVs positioning system in practice. Experimental results confirm that our proposed method outperforms the benchmark methods in terms of the positioning precision, especially under the occlusion scenarios. Moreover, ablation studies confirm the effectiveness of each unit of our method.**

***Index Terms*—Precise clustering UAVs positioning, radar-vision cross-modal, mobile spatiotemporal fusion, projective attention**

I. Introduction

Due to the advantages of flexible deployment, low cost, and equipping diverse promising techniques, unmanned aerial vehicle (UAV) clusters are widely applied in many fields, such as communication support, rescue activities and monitoring systems [1]–[3]. However, the abuse of the unauthorized clustering UAVs poses serious threats due to their extensive coverage, collaboration, and multitasking capabilities. These threats include those in physical domains (e.g., infrastructure destruction and privacy breaches) and those in information domains (e.g., communication interference and cyber attacks) [4]–[9]. Thus, it is of significant importance to monitor the clustering UAVs in order to avoid potential threat [10]–[12]. Precise real-time sensing and positioning of each UAV in the cluster are the fundamental prerequisites for achieving effective clustering UAVs monitoring.

However, positioning clustering UAVs has several unique challenges. First, since UAVs used for clustering UAVs are generally small and fly in the wide three-dimensional (3D) space, the sensed UAV features are faint, which makes it challenging to achieve precise UAV sensing and positioning [13]. Secondly, due to the high dispersal and networking capabilities of clustering UAVs, mutual occlusion among UAVs often happens, and makes it difficult to precise UAV positioning [14], [15]. Finally, the rapid flight of UAVs necessitates an effective and real-time positioning method to adapt to their fast movement [16].

Although UAV positioning systems have been developed, few of them have considered the unique challenges brought by clustering UAVs. Many existing systems focus on enhancing the sensing ability of a single modal sensor, such as visual and radar sensors, but suffer from limited sensing performance and even cannot sense clustering UAVs due to their modal properties, such as close range or coarse granularity [17]–[19]. Cross-modal UAV positioning systems are promising to achieve high positioning accuracy, but face significant challenges in modal selection and cross-modal information fusion [20]. Many of these systems are based on the fixed positioning terminals, which is difficult to locate the occluded UAVs due to the single sensing view. Some systems have used multiple positioning terminals to provide multiple views, but are limited to fixed ranges and need high deployment cost [21]. Therefore, it is imperative to overcome these limitations and realize precise real-time positioning for clustering UAVs.

In this paper, we aim at achieving accurate and real-time positioning of clustering UAVs by dynamic-view vision-radar fusion. Specifically, we exploit the long-range sensing and climate adaptability of the radar, as well as the fast and fine-grained sensing capacity of the visual camera in order to realize real-time and accurate clustering UAVs positioning. By making full use of heterogeneous modalities, the information required for sensing and positioning is more abundant and it is promising to overcome the faint feature of UAV in a single modal. In order to tackle the occlusion problem of clustering UAVs, we propose to use mobile devices as sensing terminals to perceive the cluster from multiple perspectives and mitigate errors caused by occlusion.

The research is supported in part by the National Natural Science Foundation of China(NSFC) with Grant No. 62231015, National Key R&D Program of China under Grant No. 2021ZD0110400 , Innovation Program for Quantum Science and Technology 2021ZD0302900 and ChinaNational Natural Science Foundation with No. 62132018 , 62231015. (Corresponding author: Fuhui Zhou)

Guangyu Wu and Xiang-Yang Li are with the Department of Computer Science, University of Science and Technology of China, Hefei 230052, China. (e-mail: GYWU9908@163.com; xiangyangli@ustc.edu.cn).

Fuhui Zhou is with the Key Laboratory of Dynamic Cognitive System of Electromagnetic Spectrum Space, Nanjing University of Aeronautics and Astronautics, Nanjing, China, 210016. (E-mail: zhoufuhui@ieee.org.)

Kai Kit Wong is with the Department of Electronic and Electrical Engineering, University College London, United Kingdom. (E-mail: kai-kit.wong@ucl.ac.uk.)

However, dynamic-view radar-vision clustering UAVs positioning encounters the serious spatiotemporal misalignment problem from two aspects. On one hand, disparities in the coordinate systems, sensing ranges, and sample speeds between radar and visual monitoring systems result in significant spatiotemporal mismatch. The conventional calibration approximation errors can surpass the small size of UAVs, and further result in the alignment difficulties due to the dense cluster configuration. On the other hand, the real-time motion of the clustering UAVs and the absence of fixed references result in the additional spatiotemporal misalignments in the sensed data frames. Existing reference-based data frame fusion methods have limitations due to the real-time dynamics, similar morphology of UAVs, and the absence of fixed references in the air. The exploitation of the dynamic sensing information to achieve unified UAV position information and ensure accurate positioning remains challenging.

In order to tackle those challenges, we propose a vehicle radar-vision clustering UAVs positioning system and a corresponding *a*ttention-based *s*patio*t*emporal fusion network (ASTNet) for precise clustering UAVs positioning. Our proposed positioning system is achieved by three steps, namely, the modality-oriented cross-modal feature fusion, the UAV-motion-oriented sequential cross-view feature fusion and the cross-domain positioning. To the best of our knowledge, it is the first time that a dynamic-view radar-vision clustering UAVs positioning system is developed. The main contributions of this paper are summarized as follows.

(1) We develop a novel vehicle radar-vision system for precise clustering UAVs positioning. The system exploits cross-modal sensors for precise UAV sensing and an onboard edge terminal for fast-response positioning on a patrol vehicle that provides wide-range dynamic sensing views. To address the challenges of cross-modal and cross-view information fusion, we propose a matching-based spatiotemporal fusion framework for the system. The framework novelly treats the fusion problem as an adaptive matching problem, which decreases the convention calibration challenge and the complexity. The framework adaptively matches the cross-modal features and the sequential dynamic-view features to enhance the UAV sensing areas and address the occlusion problem. Based on the enhanced UAV features, the framework precisely locates the UAVs in both the visual domain and 3D space domain through coordinate matching.

(2) We design an ASTNet that effectively realizes the proposed matching-based spatiotemporal fusion framework by using the projective attention. By abstracting the commonalities of those three tasks, the projective attention achieves the same structure to generate attention maps for enhancing clustering UAVs feature saliency. Moreover, taking into account of the distinctive characteristics of those three processes, we develop three task-specific forms of projective attention by adjusting the attention element extraction method for high effectiveness. Furthermore, we exploit modality-oriented cross-modal feature and UAV-motion-oriented cross-view UAV spatiotemporal motion to realize more effective projective-attention-based fusion for clustering UAVs positioning.

(3) We set up the system in practice and conduct actual clustering UAVs positioning experiments by using two clustering UAVs including five types of UAVs. A new radar-vision dataset with nine clustering UAVs sensing streams and over 12,000 labeled frames are collected. Compared with the benchmark methods, our framework improves the positioning precision by 18.2% at most while adapting to various challenging conditions. Meanwhile, the framework is real-time where the positioning speed achieves over 20 frames per second (FPS). The ablation studies also confirm the effectiveness of each unit of our method.

The rest of the paper is organized as follows. The related works are discussed in Section II. The vehicle-mounted radar-vision clustering UAVs positioning system is presented in Section III. In Section IV, we describe our proposed ASTNet in detail. We provide our system establishment, experimental setup, and results in Section V. Finally, we summarize this paper in Section VI.

II. Related Works

The current UAV sensing and positioning systems can be classified into two different paradigms, namely, the cooperative positioning systems and the non-cooperative positioning systems [22], [23]. In this section, we present an overview of each paradigm.

Cooperative UAV positioning systems actively associate with the UAVs during the positioning process [24]. These systems can accurately locate the UAVs in a cluster with relatively low cost based on the cooperative positioning methods. The classical systems are commonly based on UAVs sending the global navigation satellite system (GNSS) signals to the positioning terminals [25], [26]. Recently, some systems have used the UAV communication signal strength to estimate the UAV positions in GNSS-denied environments which can also achieve clustering UAVs positioning [27]. In [28], the authors explore physical characteristics of the UAVs detected in the wireless signal transmitted by UAVs during communication for positioning. These methods and systems achieve precise positioning of clustering UAVs by obtaining the position feedback information of all UAVs via mobile internet access. However, they cannot work in scenarios where UAVs do not offer active feedback, such as unauthorized clustering UAVs.

The non-cooperative UAV positioning systems are promising to address the limitations of cooperative positioning systems through external equipment without the feedback information from UAVs. Current UAV techniques are mostly single-modal sensing, including vision-based, and radar-based methods. Vision-based methods [29], [30] are intuitive but have limited range and are easily influenced by complex backgrounds. In [17], the authors proposed a spatiotemporal saliency method to enhance UAV infrared features for positioning and developed an infrared camera based positioning system. Radar-based methods [31]–[33] have long transmission ranges and high directionality but struggle with clutter interference and limited classification capacity. In [34], the authors proposed a motion-model-based UAV distinguish

method that can locate UAVs based on a low-altitude radar, and developed a positioning system by a radar. Recently, the cross-modal positioning systems have obtained higher precision. The authors proposed a modal-oriented cross-modal self-tuning fusion approach to fuse the UAV features in the visual domain and millimeter wave radar domain for obtaining a better UAV positioning performance and developed an radar-vision UAV positioning system that enhanced the UAV sensing performance in [20]. A vision-radar-spectrum collaborative method is proposed in [35], but is only tested usability in extremely close range (¡100m) and single UAV scenarios. Although these non-cooperative systems can locate multiple UAVs, they are based on fixed ground sensing terminals, which limits the positioning range, which makes it difficult to locate highly dynamic clustering UAVs.

There also exist UAV positioning systems that use the dynamic-view positioning to address UAV occlusion [36], [37]. In [36], the authors proposed a kernelized correlation filter tracker and a redetection algorithm and developed an autonomous vision-based tracking system. In [37], the authors established a Det-Fly dataset for dynamic-view UAV positioning based on an air-to-air visual detection system. However, these systems are often realized by using UAV, which have finite positioning range due to energy limitations and weak sensor carrying capacity, resulting in poor clustering UAVs positioning performance. Therefore, it is extremely challenge and important to realize a wide UAV positioning area and real-time high positioning accuracy for the multiple small UAVs even occlusion among them.

III. A Vehicle-Mounted Radar-Vision Clustering UAVs Positioning System

In this section, we first present our developed vehicle-mounted radar-vision clustering UAVs positioning system and the proposed matching-based spatiotemporal fusion framework for precise and real-time UAV positioning. Then, we define the optimization objective of our proposed framework. The main notations used in this paper are listed in Table I for easy reference.

A. System Architecture

As is shown in Fig. 1(a), we develop a vehicle-mounted radar-vision clustering UAVs positioning system to overcome the positioning challenges including dense small targets, large range, mutual occlusion, and real-time performance. Our system contains three parts (shown in Fig. 1 (b)), namely, *the patrol vehicle* that obtains the dynamic views, *the cross-modal sensors* for sensing the clustering UAVs, and *the on-board edge terminal* for positioning the clustering UAVs based on the sensed data.

1) Patrol Vehicle Offering the Dynamic Views: As highlighted in Section II, current UAV positioning systems have limitations in range and views due to the fixed sensor deployment. To overcome this, we exploit a patrol vehicle as a mobile sensing platform. The vehicle mobility allows for dynamic sensing views, expanding the sensing range with vehicle movement and eliminating the need for multiple simultaneous positioning nodes. Additionally, the patrol vehicle can adjust its position in real time based on sensing results of the clustering UAVs in order to obtain high positioning accuracy.

TABLE I
Notations used in this paper.

Symbol	Description
$\mathbf{r}$	Radar data
$\mathbf{x}$	Image data
$\mathbf{e}$	Feature
$\mathbf{p}$	Position
$t.$	Time point
T	Time interval
$\mathbf{K}, \mathbf{Q}, \mathbf{V}, \mathbf{A}$	Key, query, value and attention map
$f.$	The $(\cdot)$ function
ρ	The positioning process
$Conv.$	The convolution with kernel size of $(\cdot)$
ω	Parameters of models
$\sigma(\cdot)$	Sigmoid activation function
$\min(\cdot)$	Minimum value of a set of data
$\|\cdot\|$	The Euclidean norm of a vector
$\mathbb{E}(\cdot)$	The expectation operator
$\Pr(\cdot)$	The probability
$\mathbb{E}$	The expectation function
L	The loss function
$GeLu$	GeLU activation function [38]

2) Cross-Modal Sensors for Sensing the UAV Clusters: In order to obtain abundant features for positioning clustering UAVs, we exploit cross-modal sensors to make use of their complementary sensing information. We adopt the radar for long-range aerial scanning, which provides valuable data on the presence of airborne objects. Considering the difficulty of distinguishing UAVs from other flying objects by pulse-Doppler radar data, the system further adopts a visible light camera to sense the fine-grained appearance features of UAVs. Meanwhile, due to the real-time movement of the system, we use a GPS receiver to receive the real-time sensor position change information, so that the positioning results can be mapped into a unified coordinate system for UAVs monitoring. The details for those cross-modal sensors are presented as follows.

a) Phased Array Pulse-Doppler Radar for Long-Range Aerial Scanning: The phased array pulse-Doppler radar facilitates wide area and long-range scanning with its $n \times m$ transmit antennas. Specifically, the received baseband signal model in one coherent pulse interval Tr starting at time point tr is denoted as $\mathbf{S}_{Tr,tr}^{n\times m} \in \mathbb{C}^{n\times m}$. To reduce transmission delay caused by high-dimensional and voluminous raw data, the radar receiver preprocesses the data into target position information before transmitting it to the on-board edge terminal.

Specifically, for obtaining a high efficiency, we use the sum difference angle measurement function f_{sd} to analyze the azimuth and pitch angles of the targets, and use pulse-Doppler processing (PDP) f_{PDP} to achieve distance analysis. In order to further reduce static target interference, we use a constant false alarm rate detector (CFAR) f_{CFAR} to filter the static targets. Therefore, the received radar data $\left(\theta_{tr}^i, \gamma_{tr}^i, D^i\right)$ of the i-th target at time point tr can be expressed as

$$\begin{cases} \theta_{tr}^i, \gamma_{tr}^i = f_{sd}\left(\mathbf{S}_{Tr,tr}^{n\times m}\right) \\ D^i = \arg\limits_{c} f_{CFAR}\left(f_{PDP}\left(\mathbf{S}_{Tr,tr}^{n\times m}\right)\right) \cdot \delta \end{cases}, \tag{1}$$

This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2024.3414610

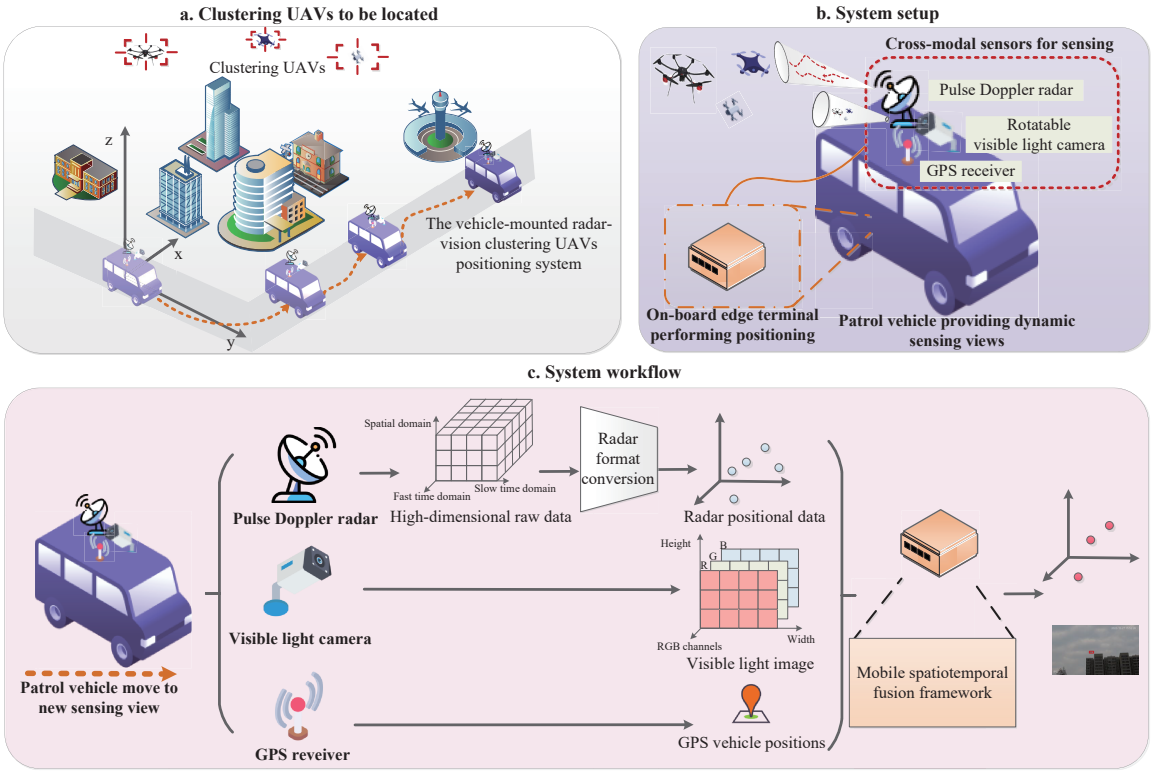

Fig. 1. The vehicle-mounted radar-vision clustering UAVs positioning system.

where θ_{tr}^i and γ_{tr}^i denote the pitch angle and the direction angle of the sensed data. D^i denotes the target distance determined by distance resolution δ and the range cell c of the target. Note that we denote all the received radar data at tr as $\mathbf{r}_{tr}$.

b) Visible Light Camera for Fine-Grained Sensing: The principle of visible light pinhole imaging imposes a constraint on the field of view angle of a visual camera, especially when using a telephoto lens for capturing images of distant clustering UAVs. A typical visible light camera with a $10\times$ zoom capability has a field of view angle of less than $10°$, which is inadequate for perceiving distant drone clusters requiring $10\times$ or greater zoom.

To overcome this limitation, our system uses a Pan Tilt to strategically rotate the visible light camera, ensuring comprehensive coverage of wide-spread clustering UAVs. Note that the visible light camera acquires red-green-blue (RGB) matrices $\mathbf{x}^t \in \mathbb{R}^{w\times h\times 3}$ at time point t, where w and h are the pixel width and height of the sensed visual image.

c) GPS Receiver for Sensor Location Recording: The GPS receiver records sensor locations, enabling the transformation of clustering UAVs positioning results into a unified coordinate system. The GPS location is denoted as $\varsigma_t = (x, y, z)$, where the initial position is defined as the origin.

3) On-Board Edge Terminal to Perform the UAV Cluster Positioning: Real-time clustering UAVs positioning is significant for effective clustering UAVs monitoring. However, many current systems require to transmit the data to the remote computation platforms. The high volume sensed data result in high transmission delay, especially when patrol vehicle faces communication instability during the large-scale patrol process. In order to overcome this issue, we propose to deploy an on-board edge terminal to perform on-board positioning, which reduces transmission delay and distance while disregarding external communication conditions. Therefore, our system achieves a higher response speed.

Furthermore, to achieve precise clustering UAVs positioning, we propose a matching-based spatiotemporal fusion framework that is conducted on the edge terminal. By leveraging the complementary nature of different sensor modalities and viewpoints, the framework can achieve precise positioning of clustering UAVs. The details of our proposed framework is presented in Section III-B.

B. Matching-Based Spatiotemporal Fusion Framework for clustering UAVs Positioning

In order to realize precise and real-time clustering UAVs positioning on the on-board edge terminal, it is important to address the spatiotemporal misalignment between cross-modal sensors and cross-view data, while simultaneously enhancing the saliency of UAV features. To bridge this gap, we propose a matching-based spatiotemporal fusion framework that has four stages, namely, data alignment and pretreatment (Sec. III-B1), cross-modal feature fusion (Sec. III-B2), cross-view feature fusion (Sec. III-B3), and multi-domain positioning (Sec. III-B4). Note that the fusion processes (Sec. III-B2, Sec. III-B3 , and Sec. III-B4) are specifically achieved by the proposed ASTNet which will be discussed in Sec. IV.

1) Data Alignment and Pretreatment: Prior to fusion and positioning, we first propose the data alignment and pretreatment unit to realize time-domain alignment of radar and visible

This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2024.3414610

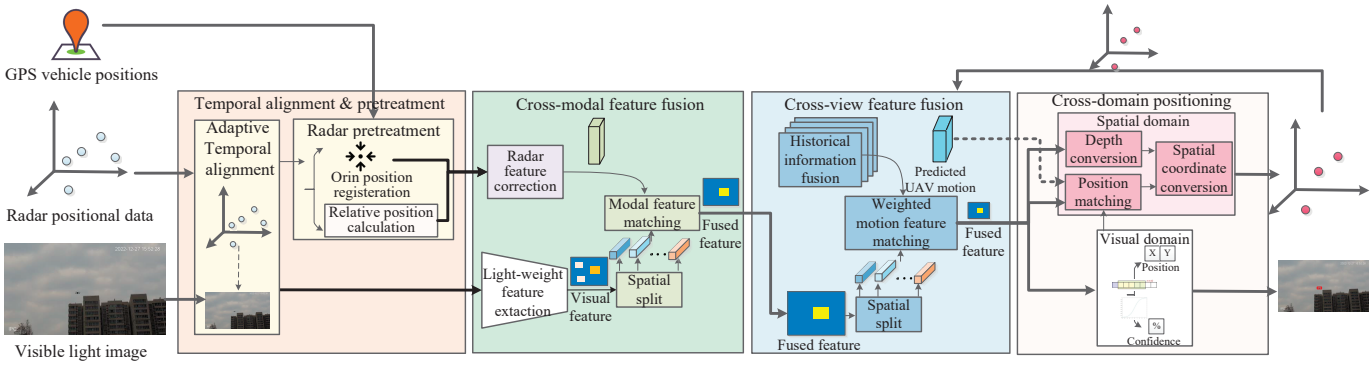

Fig. 2. The matching-based spatiotemporal fusion framework for precise clustering UAVs positioning.

light images and establish a unified coordinate system for spatial alignment.

Note that the visible light camera and the radar have severely mismatched sampling rate due to different perception mechanism, namely, the sampling rate of radar being much lower than that of camera. For example, the visible light camera can sense 25 frames per second (FPS) while the radar can require a few seconds to sense the whole range. In this case, the conventional fixed radar frame division method and the nearest neighbor alignment method result in the maximum time difference of the radar data and the visual image up to twice the radar sensing time. Therefore, we design a dynamic temporal radar-image alignment method. For each image sensed as time spot t, the corresponding radar frame contains the former radar data $\mathbf{r}^{t\sim(t-T_a)}$ within the radar sampling time T_a starting from t. By performing our temporal radar-image alignment method, each visual image is matched with its nearest radar full range sensing result, with a maximum time difference of inevitable radar perception time.

Considering the dynamic nature of the positioning process, it is desirable to have a unified coordinate system for the three-dimensional representation of the positioning results in order to facilitate the fusion of subsequent modes and perspectives. To achieve this purpose, we unify the three-dimensional coordinates perceived by the radar into a consistent coordinate system. This coordinate system is established with the patrol vehicle coordinates at the start time t_0 as the origin, and all subsequent radar coordinates are converted from their perceived position coordinates to this unified coordinate system. The radar perception coordinates are transformed by

$$\mathbf{r}_t = \mathbf{r}_t - (\varsigma_t - \varsigma_{t_0}), \tag{2}$$

where ς_{t0} is the original patrol vehicle position and defined as the origin. Due to the fixed orientation of the radar, the perception coordinate system at each moment is a translation relationship with the same coordinate system. Therefore, subtracting the relative position relationship $(\varsigma_t - \varsigma_{t_0})$ can achieve coordinate system conversion.

2) Cross-Modal Feature Fusion: The cross-modal feature fusion unit aims at accurate clustering UAVs sensing by extracting the radar data features and the visual image feature and further conducting feature fusion. However, due to calibration difficulty brought by the difficulty in distinguishing targets types of radar and the real-time imaging parameter adjustment of the camera, conventional calibration-based methods can hardly achieve accurate calibration and effective modal fusion. Therefore, we propose a cross-modal feature fusion unit to address this challenge by replacing calibration through cross-modal feature matching.

First, we separately extract radar and visual features to achieve feature-level fusion for higher fusion effectiveness. Specifically, the extracted radar feature is denoted as $\mathbf{e}_r \in \mathbb{R}^{1\times NUM_{tar}}$ and the extracted visual feature is denoted as $\mathbf{e}_x \in \mathbb{R}^{w_e\times h_e\times c_e}$, where NUM_{tar} denotes the radar target number. Each target position is embedded into one feature, and w_e, h_e, c_e denote the width, height and channel number, respectively.

Meanwhile, since we aim at effective feature matching for the fusion, we split the visual features from the space aspect. Specifically, the visual feature is splited into $\hat{\mathbf{e}}_x \in \mathbb{R}^{NUM_A=(w_e\times h_e)\times c_e}$.

After splitting the visual feature, we match the cross-modal features to fuse the visual and radar features, given as

$$\mathbf{e}_{mpa} = f_{mm}(\mathbf{e}_r, \hat{\mathbf{e}}_x), \tag{3}$$

where $\mathbf{e}_{mpa}$ denotes the fusion result, f_{mm} denotes the matching function which is further realized by the cross-modal projective attention in Sec. IV.B.

3) Cross-View Feature Fusion: Based on the fused cross-modal features, the framework exploits the cross-view fusion unit to realize effective UAV sensing information complementarity between dynamic views. However, many methods only fuse information from adjacent data frames, making it difficult to achieve effective perception and position prediction for UAVs with long-term occlusion. Therefore, we propose a cross-view information fusion method f_h that achieves the fusion of all historical positioning result information $\{\mathbf{p}_{3D}{}^{t_0}, \mathbf{p}_{3D}{}^{t_1}, \cdots \mathbf{p}_{3D}{}^{t}\}$ and further estimates the current UAV positions, given as

$$\ddot{\mathbf{e}}_{cpa} = f_h\left(\mathbf{p}_{3D}{}^{t_0}, \mathbf{p}_{3D}{}^{t_1}, \cdots \mathbf{p}_{3D}{}^{t}\right), \tag{4}$$

where $\ddot{\mathbf{e}}_{cpa}$ denotes the fusion result, which outputs the estimated UAV positions at the current time point. Specifically, f_h is realized by Eq. (15).

Meanwhile, considering the spatiotemporal misalignment between each view, it is difficult to fuse the historical information with the current information. Similiar to the cross-modal feature fusion, we view the cross-view feature fusion with a matching problem. The estimated UAV positions are achieved

by the historical positioning result matching with the cross-modal feature, given as

$$\mathbf{e}_{cpa} = f_{vm}\left(\mathbf{e}_{mpa}, \ddot{\mathbf{e}}_{cpa}\right), \tag{5}$$

where $\mathbf{e}_{cpa}$ denotes the fusion result, f_{vm} denotes the matching function, which is further realized by the cross-view projective attention in Sec. IV.C.

4) Cross-Domain Positioning: The cross-domain positioning unit achieves both the pixel-level positioning and the 3D positioning based on the cross-view features. In this case, the positioning results can provide good visualization at the pixel level while providing 3D coordinate information for cross-view fusion.

Specifically, the pixel-level positioning calculates the positions $\mathbf{p}_{2D}$ of the targets and the corresponding confidence $\Pr(\mathbf{p}_{2D})$. Then, the pixel-level positions are converted into the established 3D coordinate to provide UAV motion for cross-view fusion. Specifically, the depth of each pixel $\mathbf{z}$ can be estimated based on the cross-view feature. Based on the pixel depth and pixel-level UAV positions, their 3D positions $\mathbf{p}_{3D}$ can be achieved based on a reverse perspective transformation, expressed as

$$\left(\mathbf{p}_{3D}, \mathbf{1}\right)^{\top} = \mathbf{z}\mathbf{\Gamma}\left(\left(\mathbf{p}_{2D}, \mathbf{1}\right)^{\top}\right), \tag{6}$$

where $\mathbf{\Gamma}$ is the transformation matrix. Note that the acquisition of $\mathbf{z}$ and $\mathbf{\Gamma}$ will be further discussed in Sec. IV. D.

C. Objective Definition

To precisely locate each UAV in the clustering UAV, the objective is to minimize the positioning error of the perceptible UAVs and maximize the prediction precision of occluded UAVs in the occluded area $\mathbf{O}$. Thus, we define the objective of the proposed framework as

$$\min \sum_{i=1}^{|\mathbf{p}|} \|\mathbf{p}_i - \mathbf{p}_i^*\| + \sum_{j=1}^{|O|} |cnt(\mathbf{p}^o \subseteq O_j) - cnt(\mathbf{p}^{o*} \subseteq O_j)|, \tag{7}$$

where $\mathbf{p}_i$ denotes the predicted position of the i-th UAV in the sensed data and $\mathbf{p}_i^*$ denotes its ground-truth position. O_j is the j-th area in the sensed data that contains the occluded UAVs. $\mathbf{p}^o$ is the predicted UAV position in the occluded areas while $\mathbf{p}^{o*}$ denotes the ground-truth positions of the occluded UAVs. $cnt(\cdot)$ denotes the counting function.

Furthermore, we divide the objective of the clustering UAVs positioning into two sub-objectives for ease optimization, namely, precisely outputting the UAV positions in the visual image domain and the established three-dimensional unified coordinate system. Those objectives are given in Eq. (8) where $\rho_v\left(\cdot;\omega_{\rho v}\right)$ and $\rho_r\left(\cdot;\omega_{\rho r}\right)$ denote the positioning process of the visual image coordinate domain and the unified coordinate system, respectively. $\omega_{\rho v}$ and $\omega_{\rho r}$ are their parameters, respectively. For the image domain UAV positions output in Eq. (8a), the objective aims at optimizing the parameter $\omega_{\rho r}$ of the positioning process to gain the accurate UAV output positions $\rho_v\left(\widehat{\mathbf{e}};\omega_{\rho v}\right)[0]$, with high corresponding confidence $\rho_v\left(\widehat{\mathbf{e}};\omega_{\rho v}\right)[1]$. For the 3D UAV positions output in Eq. (8b), since this process is without manual labels, the objective aims at maximizing its gain on the pixel level positioning results at the next moment to achieve self-supervision.

IV. ASTNet for Matching-Based Spatiotemporal Fusion Framework

As is shown in Fig. 3, we present the ASTNet to perform precise feature fusion and clustering UAVs positioning. Based on our developed system, the ASTNet consists of a modality-oriented cross-modal feature fusion and a UAV-motion-oriented cross-view fusion to effectively fuse the sparse radar features with the dense visual features, and a cross-domain UAV positioning for the pixel-level positioning on the visual image and the 3D UAV positioning in the unified coordinate.

A. Projective Attention Structure for Matching-Based Fusion

Effectively matching heterogeneous features is the basis of the effectiveness of the ASTNet. Thus, we treat the matching process as an attention process and introduce projective attention, leveraging the attention mechanism to uncover feature correlations. The attention mechanism has three components, namely, the query, the key and the value. The query calculates correlations with the key to generate an attention map, which is then fused with the value to enhance value saliency. However, the conventional self-attention mechanisms cannot be applied to to match and fusion the heterogeneous features since they realize all three components based on the same feature.

In order to tackle the limitation, the projective attention incorporates heterogeneous features for the query and key. Considering the common characteristics within the cross-modal feature fusion, cross-view feature fusion and cross-domain feature fusion, which match and fusion the highly directional positional features and classifiable features, our projective attention uses the positional features as the query $\mathbf{Q}$, while using the classifiable features as the key $\mathbf{K}$. Moreover, in order to generate the attention map $\mathbf{A}$, our projective attention matches each element in the key with all the query elements, realizing more effective correlation matching, given as

$$\mathbf{A} = \sigma\left(\mathbf{Q} \times \left(reshape\left(\mathbf{K}\right) \otimes \mathbf{1}_{\mathbf{Q}}\right)\right), \tag{9}$$

where we reshape the key into 1D and expand each element into the query element amount based on function $reshape$. In this case, all the key elements can be matched with the query and acquire the matching result by matrix multiplication. In projective attention, each element is independent after being matched rather than cross-correlated as the existing attention mechanisms. Therefore, unlike existing attention mechanisms that use the cross-correlation-based softmax function for attention allocation, we use sigmoid function σ to independently normalize each element and generate the attention map.

Upon achieving the attention map $\mathbf{A}$, the projective attentions aims at enhancing the value $\mathbf{V}$ based on $\mathbf{A}$. Specifically, in order to acquire a more accurate positioning result, we define the $\mathbf{V}$ as the classifiable features. Considering each element in $\mathbf{A}$ can correspond to that in $\mathbf{V}$, our projective attention achieves the enhancement based on the dot multiplication, given as

$$\mathbf{e}_{pa} = \mathbf{A} \cdot \mathbf{V}, \tag{10}$$

where $\mathbf{e}_{pa}$ denotes the projective attention result. In this context, each fusion unit aims at achieving the appropriate key, query and value for the projective attention.

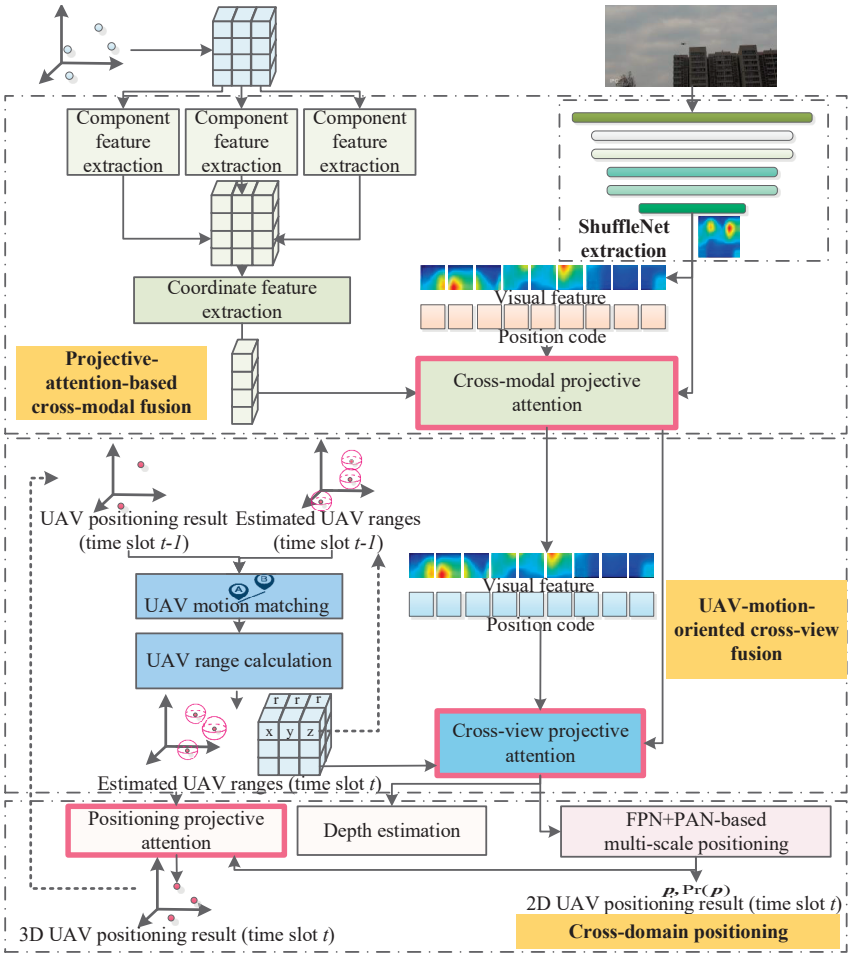

Fig. 3. Flowchart of the ASTNet.

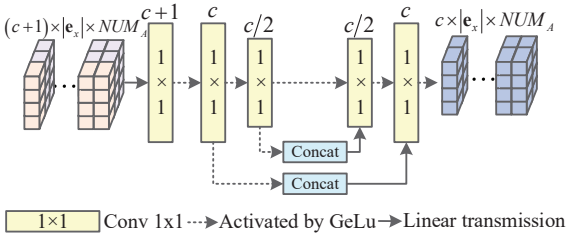

Fig. 5. The channel encoding block structure.

B. Modality-Oriented Cross-Modal Feature Fusion

Conventional calibration-based fusion methods have limitations in two aspects. On one hand, the small UAV size amplifies the radar range and angle deviations, which requires a more precise fusion approach. On the other hand, the fast moving clustering UAVs necessitate real-time adjustment of internal and external parameters, which results in the fixed calibration parameters ineffective. In this case, the conventional methods can hardly achieve the accurate fusion for clustering UAVs positioning. In contrast, we aim to realize the modality-oriented cross-modal feature fusion.

First, to reduce the modality inference for accurate fusion, we extract less-interfered semantic features for each modality based on its specific characteristics. For radar data, we address two types of interference. On one hand, the phase errors in the pulse echo can cause coordinate deviation of the sensed data. To reduce this deviation, we propose the radar component feature extraction block which extracts relative coordinate features from three dimensions of radar data by an MLP with weight sharing. The process is given as

$$\mathbf{e}_r = \begin{bmatrix} M_x\left(\hat{\mathbf{r}}[:,0,:::]\right); M_y\left(\hat{\mathbf{r}}[:,1,:,:]\right); \\ M_z\left(\hat{\mathbf{r}}[:,2,:::]\right) \end{bmatrix}_{\sigma}, \tag{11}$$

where $\mathbf{e}_r$ denotes the extraction result. M_x, M_y and M_z denote the MLPs for x, y, and z component, respectively, and are all activated by the sigmoid function σ. In this case, extracting features from three dimensions separately correct pitch and azimuth errors, and the weight sharing MLP correct distance errors. On the other hand, each radar-sensed coordinate exists classification disability (i.e., not necessarily a UAV position). In this case, we propose the coordinate feature extraction module to extract each radar-sensed 3D coordinate into 1D feature for ease of fusion with the visual features. The process is given as

$$\mathbf{e}_{\hat{r}}[i,:,:,:,:] = \sigma\left(Conv^{i}_{3\times 1}\left(\mathbf{e}_r[i,:,:,:,:]\right)\right), \tag{12}$$

where $\mathbf{e}_{\hat{r}}$ denotes the extracted result that is achieved by the 3×1 convolution at each sensed position.

For the visual images, the sensed UAVs have small areas and the backgrounds are complex. Therefore, we use the lightweight ShuffleNet to extract UAV visual feature which avoids excessive downsampling to preserve the small UAV

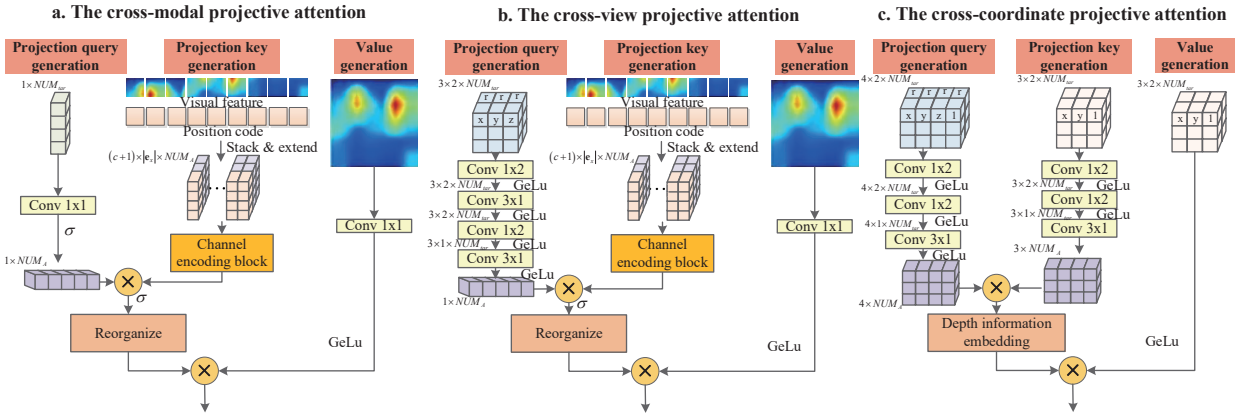

Fig. 4. The structures of projective attention for cross-modal, cross-view and cross-coordinate situations.

$$\min_{\omega_{\rho v}{}^t}(\mathbb{E}\left(\rho_v\left(\widehat{\mathbf{e}};\omega_{\rho v}\right)[0]-\mathbf{p}_{2D}\right)+\mathbb{E}\left(\rho_v(\widehat{\mathbf{e}};\omega_{\rho v})[1]-\Pr(\mathbf{p}_{2D})\right))^t, \tag{8a}$$

$$\arg\min_{\omega_{\rho r}{}^t}\left(\mathbb{E}\left(\rho_v\left(\widehat{\mathbf{e}};\omega_{\rho v}\right)[0]-\mathbf{p}_{2D}\right)+\mathbb{E}(\rho_v(\widehat{\mathbf{e}};\omega_{\rho v})[1]-\Pr(\mathbf{p}_{2D}))^{t+1}\left|\rho_r\left(\rho_v\left(\widehat{\mathbf{e}};\omega_{\rho v}\right);\omega_{\rho r}\right)^t\right.\right), \tag{8b}$$

areas while downsampling complex background into insignificant features.

Then, we propose a cross-modal projective attention based on the projective attention structure to match the cross-modal features and achieve the fusion. As is shown in Fig. 6 (a), we exploit the directionality of radar features as the attention query $\mathbf{Q}_m$ and the fine-grained visual features as the key $\mathbf{K}_m$ in order to enhance the UAV feature in the visual domain. Specifically, we use a fully connected layer to reorganize the radar features as $\mathbf{Q}_m$, denoted as

$$\mathbf{Q}_m = Conv_{1\times1}\left(\mathbf{e}_{\widehat{\mathbf{r}}}\right). \tag{13}$$

For the visual features, we first downsample the visual features into single channel. Then, we flatten the visual features into one dimensional and expand each feature element into the radar feature number. Then, we design a channel encoding block (CEBLock) f_{ceb} to embed location and feature values and achieve the key generation. The CEBlock uses a residual-connected encoder decoder to encode position information into each feature channel and ultimately achieve key generation, given as

$$\mathbf{K}_m = f_{ceb}\left(Flatten\left(\left[\mathbf{e}_x; p(\mathbf{e}_x)\right]\right)\right), \tag{14}$$

where $\mathbf{K}_m$ denotes the generated key. $Flatten\left(\left[\mathbf{e}_x; p(\mathbf{e}_x)\right]\right)$ is the flatten process of the stacked feature-position pair $[\mathbf{e}_x; p(\mathbf{e}_x)]$. Meanwhile, the attention value $\mathbf{V}_m$ is also generated based on the visual feature with 1×1 convolution. In this case, the cross-modal projective attention components are achieved and the cross-modal fusion result $\mathbf{e}_{mpa}$ is achieved based on Eq. (9) and Eq. (10).

C. UAV-Motion-Oriented Cross-View Fusion

As discussed in Section III, we exploit UAV positioning results from dynamic views to achieve more precise positioning of UAVs in cluster and accurate prediction of occluded UAVs. However, due to the variable motion trajectories of the UAVs and the real-time movement of the patrol vehicle, effective correlation and information fusion between front and rear perspectives remain extremely challenging.

To address this challenge, we exploit the UAV motion characteristics, namely, the *3D flight* and *upper flight speed limit* v. In this case, each UAV has a fixed 3D motion area at time point t given the sampling interval T_s (shown in Fig. 6(a)), expressed by Eq. 15,

where $\mathbf{p}_{3D}^{t,i}$ is the position for the i-th UAV at the current time point t, the estimated motion area is a spherical region with radius $T_s \cdot v$ and center $\mathbf{p}_{3D}^{t,i}$. Then, considering the presence of occlusion and missed positioning, we use historical information for UAV motion area estimation. Specifically, for the predicted area $\mathbf{P}_m^t$, if there are no UAVs in that area at the next time interval $t+1$, the area will be retained and the radius will increase $T_s \cdot v$, which will be added to the estimated results at the time interval $t+2$.

While the estimated motion areas represent the cross-view information, we then focus on fusing the cross-view information with the cross-modal feature information $\mathbf{e}_{mpa}$. Note that the cross-view fusing problem resembles the cross-modal fusion problem, where the 3D cross-view UAV coordinates need to match and enhance the 2D cross-modal UAV features. In light of this, we propose a cross-view projective attention as is shown in Fig. 4(b).

Unlike cross-modal fusion, the cross-view fusion process involves UAV range information, resulting in more complex UAV position representation. For this reason, we design a range-involved query generation for range information of UAVs for cross view attention. Specifically, we first use 1×1 convolution to embed range information for each coordinate component, then use 3×1 convolution to achieve fusion expression for each coordinate, and finally use 1×1 convolution to expand the coordinates to the visual feature space for subsequent alignment. The query generation result is denoted as $\mathbf{Q}_v$. Meanwhile, the attention key $\mathbf{K}_v$ is generated by a CEBlock based on $\mathbf{e}_{mpa}$. Furthermore, we use a 1×1

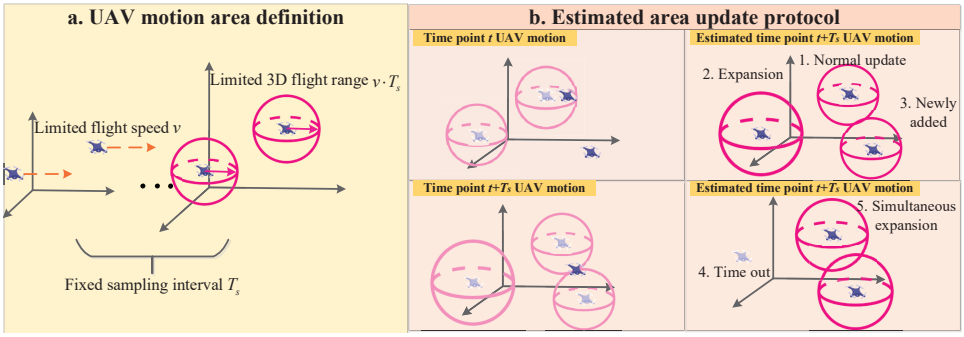

Fig. 6. The UAV motion characteristics.

$$\mathbf{p}_{3D}^{t,i} \in \left\{ (x,y,z) \,|\, \left(x - \mathbf{p}_{3D}^{t',i}[0]\right)^2 + \left(y - \mathbf{p}_{3D}^{t',i}[1]\right)^2 + \left(z - \mathbf{p}_{3D}^{t',i}[2]\right)^2 \leqslant (t-t') \cdot T_s \cdot v \right\}, \tag{15}$$

convolution layer activated by GELU function to recombine the fused cross-modal features as the attention value $\mathbf{V}_v$. In this case, the cross-view projective attention can be achieved based on Eq. (9) and Eq. (10), where $\mathbf{e}_{vpa}$ denotes the fused dynamic-view feature by the cross-view projective attention.

D. Cross-Domain Positioning

Corresponding to the framework, the cross-domain positioning of ASTNet aims to achieve the 2D and 3D positions of the sensed clustering UAVs. First, to achieve higher 2D positioning accuracy and efficiency for various UAVs in the cluster, a feature pyramid network (FPN) combined with a pixel aggregation network (PAN) is exploited. Based on the positioning results of three scales, the non maximum suppression (NMS) is exploited to exclude invalid positions and decide the final UAV positions $\mathbf{p}_{2D}$ and the corresponding confidence $\Pr(\mathbf{p}_{2D})$. The details can be found in [39].

Based on the positioning results in the visual domain, the cross-domain positioning of ASTNet represents the 3D UAV positions P_{3D} within the defined unified coordinate system. Due to variations in camera parameters and depth uncertainty, direct application of the reverse perspective transformation is challenging. Therefore, we propose a cross-coordinate projective attention (shown in Fig. 4(c)) to adaptively realize the transformation by exploring input relationships based on attention mechanism.

Specifically, the cross-coordinate projective attention first realizes the transmission matrix $\mathbf{\Gamma}$ by aligning the estimated UAV motion (i.e., projection query) and the visual positioning results (i.e., projection key). Then, the result is multiplied with the depth value $\mathbf{z}$ estimated by Fastdepth method [40]. In this case depth information embedding is achieved and the attention map is generated. Finally, based on Eq. (6), the 3D UAV positions P_{3D} is achieved by multiplying the attention map with the visual positioning results.

E. Algorithm Update

Different from the conventional DNN methods which provide labels for all modalities, radar labels are absent in our work considering the difficulty in distinguishing between target information and interference information in actual scenarios. Thus, we consider an end-to-end training strategy, which directly uses image labels to provide supervision for 2D visual domain positioning procedure during the training process, and achieves self-supervision for the 3D positioning.

For the 2D positioning, the GIoU loss is used for 2D positioning, denoted as L_p. Then, in order to improve the feature accuracy, the regression loss is achieved by calculating the cross entropy between the predicted and the ground-truth confidence, denoted as L_c.

In order to improve the cross-modal, cross-view, and cross-domain feature fusion accuracy and improve the 3D positioning capability, we define the spatiotemporal fusion loss L_s. Specifically, we generate a ternary mask that denotes the desired fused feature structure and use the structural similarity $SSIM$ to achieve L_s [41], given as

$$L_s = \frac{1}{2} SSIM(\mathbf{e}_{mpa}, M^{label}) + \frac{1}{2} SSIM(\mathbf{e}_{vpa}, M^{label}), \tag{16}$$

where M^{label} denotes the ternary mask generated by the visual UAV lables. The mask value for areas without UAVs is 0, the mask value for areas with a single UAV is 1, and the mask value for areas with multiple overlapping UAVs is the number of UAVs. Specifically, the spatiotemporal fusion loss L_s is the summarize of the cross-modal structural similarity and cross-view feature structural similarity.

The overall loss function to update the proposed algorithm is achieved by aligning the above loss functions, given as

$$L = \lambda_1 L_p + \lambda_2 L_c + \lambda_3 L_s, \tag{17}$$

where λ_1 ,λ_2 and λ_3 denote the weight of the loss functions. Then, the backpropagation using gradient descent is achieved based on L to train and update the proposed method.

V. Performance Evaluation

In this section, we first demonstrate the actually built vehicle-mounted clustering UAVs positioning platform and the corresponding testing environment. Then, experimental results are presented to compare the performance of our proposed ASTNet with that of other benchmark methods based on our established platform. Furthermore, The ablation study is also

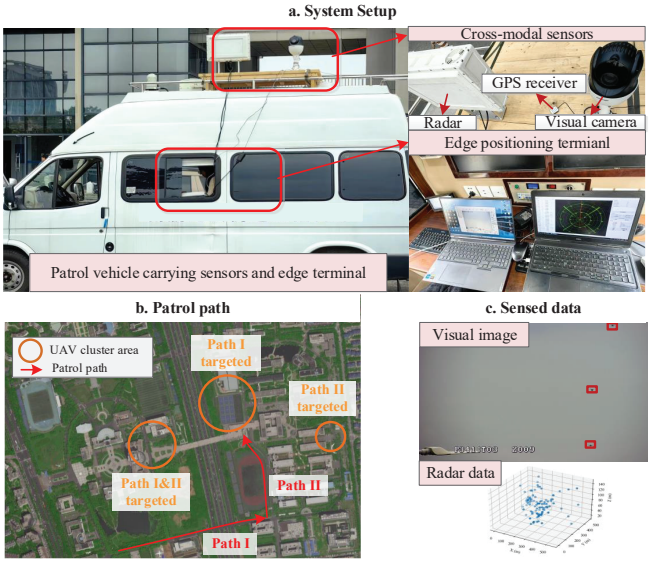

Fig. 7. The actual vehicle-mounted clustering UAVs positioning system.

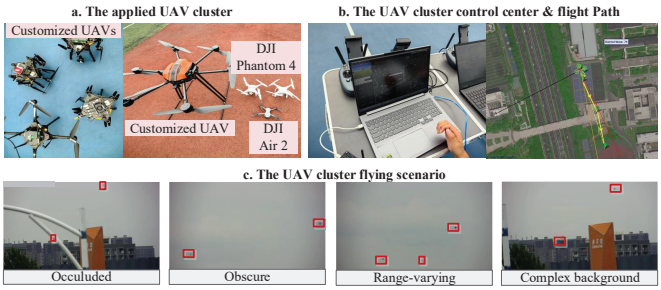

Fig. 8. The clustering UAV platform providing actual positioning scenario.

TABLE II
THE DATA PROPORTION OF EACH CHALLENGING CONDITION.

	Occluded	Obscure	Dense	Complex background
Propotion	14.3%	23.6%	28.6%	19.1%

demonstrated to confirm the effectiveness of each unit in the ASTNet.

A. Vehicle-mounted Clustering UAV Positioning Platform

As is shown in Fig. 7 (a), we establish our vehicle-mounted clustering UAVs positioning system for clustering UAVs positioning task. We add a sensor deployment bracket and Pan Tilt at the top of a patrol van to enable cross modal sensor installation. At the same time, we modify the internal space of the van to include an edge terminal based on NVIDIA RTX3060 and connect sensors. Meanwhile, our system can be easily deployed on other vehicles. The visual camera, the phased array pulse Doppler radar, and the GPS receiver are deployed in a fixed relative position on the bracket. The parameters of the sensors are summarized in Table III. Furthermore, we achieve the positioning of clustering UAVs in three distinct airspace based on two patrol routes (shown in Fig. 7 (c)), where the patrol coverage area exceeds 1 km^2 and the clustering UAV.

Meanwhile, as is shown in Fig. 8, we establish a clustering UAV platform to provide an actual positioning scenario for evaluation. Specifically, we adopt four types of UAVs with different sizes, including DJI Phantom 4, DJI Air 2 and two types of customized UAVs, as is shown in Fig. 8 (a).

Based on this platform, the clustering UAVs positioning data are collected, including 10 positioning data streams with the total duration over 2.1 hours, where the distances between the clustering UAVs and the patrol van various from 300m to 800m. Meanwhile, the patrol van collects five positioning data streams are collected on each patrol path in Fig. 7. We select 12,000 data frames to form the dataset. We present some sample scenarios in Fig. 8 (c). It is clear that our collected data contains typical clustering UAVs positioning challenges, including occlusion, obscure, dense and complex background. We have summarized the proportion of data for each challenging conditions in Tab. II.

Specifically, the dataset has two types of divisions. For positioning performance evaluation, we divide the dataset into training set and testing set at the proportion of 80% and 20%. Note that the data are randomly sampled from each sequence, while we also guarantee that the each challenging scenario has similar data distribution for ease of evaluation. For generalization evaluation, we separately divide the data collected for path I and path II into training and testing sets (denoted as I and II). For evaluation, we label all UAV positions in all images while the radar data are unlabeled. We provide two types of labeling results, namely labeled the occluded UAVs (Occlude) and not labeled the occluded UAVs (Non-occlude). Meanwhile, considering ASTNet requires s-patiotemporal fusion, we provide each data with positioning results at the last time point during the training and testing process.

B. Experiment Settings and Performance Metrics

We train the proposed method along with baseline methods for 200 episodes on the dataset. Moreover, optimizers of three networks are all set as Adam [42]. The ShuffleNet backbone follows the settings used in [43]. Three Res3DBlocks and four feature merging blocks are exploited in the MOSSNet.

We compare our methods with both the single-modal methods and radar-vision fusion methods. Five vision-based methods are adopted [43]–[45], including a two-stage positioning method, a one-stage positioning method, a light-weight method, a NAS-based method and an attention-based method. Two radar data positioning methods are adopted, including two CFAR-based methods [46], [47]. Three fusion methods, CRFNet [48], FusionNet [49], and SFFR [50] are adopted to evaluate the effectiveness of ASTNet in feature fusion, where CRFNet and FusionNet are feature-level fusion and SSFR is radar point-level fusion.

Five performance metrics are adopted to evaluate the UAV positioning quality from different aspects, including mean average precision (mAP), precision, recall, frame per second (FPS) and intersection over union (IoU) [51]. mAP, precision and recall focus on evaluating positioning precision. FPS shows the efficiency of the method while IoU indicates the position accuracy. Moreover, we define that the positioning result is regarded as a UAV when its IoU with the groundtruth is above 0.6.

TABLE III
SENSOR CONFIGURATIONS OF THE ESTABLISH PLATFORM.

Visual camera parameter	Value	Radar parameter	Value	GPS receiver	Value
Type	P20A1POE	Type	D6000	Type	-
Resolution	1920×1080	Frequency	15.6 GHz	Frame rate	10FPS
Frame rate	25 FPS	Frame rate	$\frac{1}{6}$FPS		
Zoom factor	1-23				

TABLE IV
RESULTS ON THE REAL-WORLD DATASET USING OUR METHOD AND METHODS BASED ON SINGLE MODALITY. NON-OCCLUDE: NOT LABELING THE OCCLUDED UAVS. OCCLUDE: LABELING THE OCCLUDED UAVS.

	Methods	Precision		Recall		mAP		Speed
		Non-occlude	Occlude	Non-occlude	Occlude	Non-occlude	Occlude	(FPS)
Vision	YOLOv5	91.5%	67.6%	73.1%	42.4%	72.1%	67.4%	31.4
	Faster-RCNN	95%	66.1%	74%	51.3%	75.2%	68.5%	18.6
	YOLOv5-lite	94.8%	69%	70.2%	35.9%	72.6%	66.6%	36.1
	YOLO-NAS	94.2%	68.7%	73.7%	53.3%	74.5%	64.8%	14.7
	RT-DERTR	97.4%	69.7%	72.9%	66.2%	77.3%	69.7%	10.3
Radar	CA-CFAR+DBSCAN	3.0%	3.2%	63.7%	62.0%	-	-	5.7
	GO-CFAR+DBSCAN	3.5%	3.3%	64.6%	61.7%	-	-	5.9
Vision+radar	Proposed ASTNet	93.8%	93.0%	85.2%	83.7%	89.8%	88.7%	20.2

TABLE V
RESULTS ON THE REAL-WORLD DATASET OBTAINED WITH OUR METHOD AND BASELINE RADAR-VISION-FUSION METHODS. NON-OCCLUDE: NOT LABELING THE OCCLUDED UAVS. OCCLUDE: LABELING THE OCCLUDED UAVS.

	Calibration	Precision		Recall		mAP		Speed (FPS)
		Non-occlude	Occlude	Non-occlude	Occlude	Single	Occlude	
CRFNet	✓	60.2 %	53.9%	54.6%	53.4%	50.1%	49.7%	18.2
FusionNet	✓	58.1%	56.7%	51.1%	52.1%	50.4%	48.2%	16.8
SSFR	✓	57.3%	50.5%	71.1%	62.3%	58.2%	53.1%	18.7
Proposed ASTNet	✗	93.8%	93.0%	85.2%	83.7%	89.8%	88.7%	20.2

C. Performance Comparison

1) General Positioning Performance: We first present the general positioning performance comparisons of different methods, including the single-modal methods and the radar-vision methods. Specifically, we achieve the comparison under two types of labeling methods: labeling the occluded UAVs and not labeling the occluded UAVs, in order to illustrate the clustering effects of the UAVs to the positioning methods.

Comparing with the single modal methods. Tab. IV shows the performance comparison among different single-modality methods. Especially, we provide two types of labeling methods, namely, labeling the occluded UAVs and not labeling the occluded UAVs, in order to better demonstrate the effectiveness of different methods. It is seen that the positioning performance, especially in the occluded scenario, has consistent improvement when our method is used, indicating that the exploitation of cross-modal and cross-view data benefits for precise positioning. Specifically, the vision-based methods can drop 6.7% mAP at most when labeling the occluded data, the proposed ASTNet only drops 0.7%, due to the usage of radar and the cross-view information. When cross-view information is absent, the labeled occluded UAVs can be interference for training the visual methods, since the occluded area provides false UAV features. We further observe that the radar-based positioning methods generally have lower precision compared with the image-based methods, due to their poor classification ability. Meanwhile, the radar-based methods have generally high recall due to its ability to locate moving objects.

Comparing with the Radar-Vision Fusion Methods.

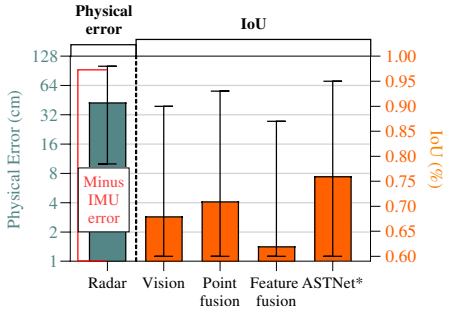

Fig. 9. Positioning range error of different methods.

Tab. V verifies the performance of positioning UAVs by our method compared to other radar-vision fusion based methods. The calibration-based methods shows very low positioning accuracy since they are based on fixed calibration parameters. In our scenario, due to the changing views and rotating camera, the fixed calibration parameters are not suitable. Meanwhile, we also observe that the cross-modal methods show similar performance to the occlusion scenario, which confirms the effectiveness of modal complementarity. Since our method is achieved based on the adaptive matching, our method can adapt to the dynamic views. In the meanwhile, due to the self-tuning cross-modal affine transfer, the feature saliency is significantly enhanced while the calibration process is omitted, which further saves the positioning time compared with the calibration-based methods.

Positioning Range Error. Moreover, we present the positioning range error of different methods based on IoU metric. As is shown in Fig. 9, we can observe that the vision-based

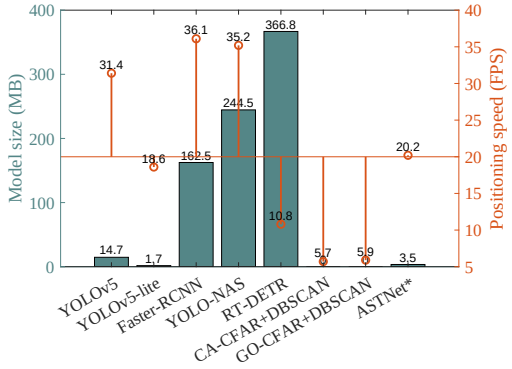

Fig. 10. Complexity comparison based on the inference speed and model size. 20FPS is the real-time criterion.

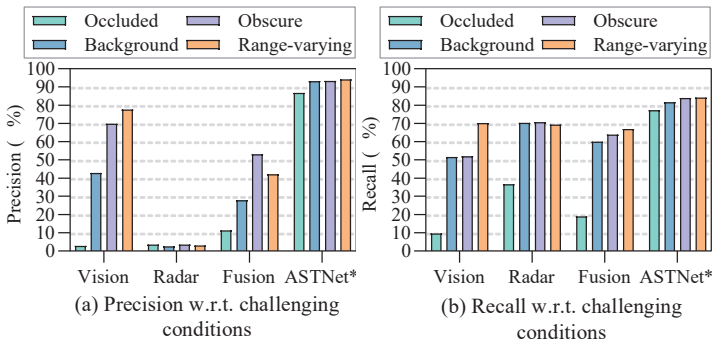

Fig. 12. The positioning accuracy of different methods in challenging conditions.

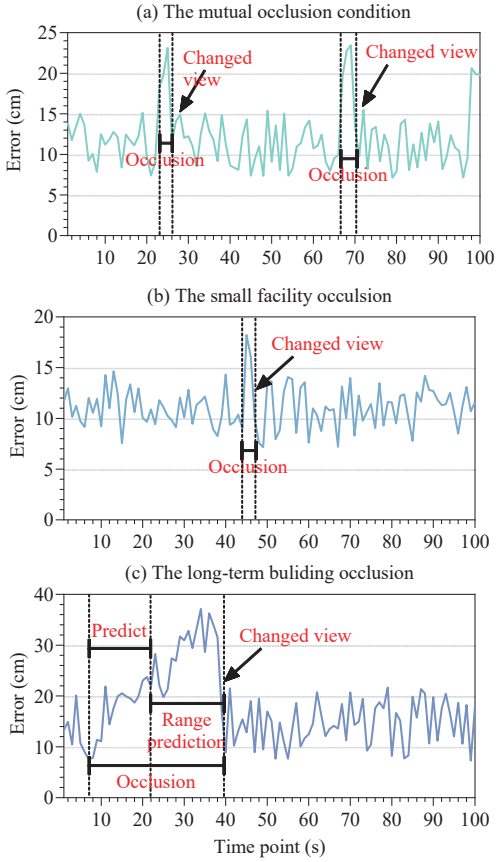

Fig. 13. The positioning error of ASTNet under different occlusion conditions. We transform IoU into the maximum distance between the UAV area and ground-truth center, since the ASTNet provides UAV motion area in long-term occlusion.

approaches have higher IoU compared to the current radar-vision-based approaches, since the calibration error and the radar sensing error enlarges the positioning error of radar-vision-based approaches. Our methods can reduce the error from two aspects. First, we realize the projective-attention-based cross-modal fusion, which is a soft weighting method based on cross modal information instead of existing hard supervision methods using radar. Meanwhile, we realize the UAV motion area prediction based on the spatiotemporal fusion, realizing more salient UAV features for positioning. For the radar-based methods, we provide their physical positioning range error with the UAVs based on the UAV IMU information acquired at the UAV controlling end, where IMU positioning error is considered. It is obvious that radar-based methods have large positioning range error, which can exceed the UAV size. This is due to the presence of direction finding errors and spatial signal interference in the radar, resulting in inaccurate direction finding and significant positioning errors. Such error and the small size characteristic of UAVs decrease the efficiency of current radar-vision methods.

Complexity Comparisons. The model complexity and real-time performance are compared in Fig. 10. We define 20FPS as the threshold of the real-time performance. It shows that our method obtains relatively high model size, since we require the three proposed fusion processes to achieve higher positioning performance. Since our method can simultaneously process the radar and image data, we still achieve the real-time positioning while the image-based Faster-RCNN, YOLO-NAS, RT-DETR and the radar-based CFAR methods cannot achieve. Since information content that a single modality can provide is limited, simply mining single mode information can lead to a surge of parameters, while the effect on the task is not as good as the introduction of new modes. This causes the unsatisfied real-time performance of Faster-RCNN. The slow inference speed of CFAR+DBSCAN methods is because CFAR and clustering of high-dimensional radar data are slow and require original radar data transmission.

2) Positioning Accuracy in Challenging Conditions: We further compare the positioning performance of different methods in the challenging conditions as is shown in Fig. 8. Specifically, we have presented the visualization results of the proposed ASTNet in these conditions as is shown in Fig. 11. It shows that our proposed ASTNet can achieve effective positioning in these conditions. Furthermore, since the challenging conditions do not affect the complexity of different methods, we give the numerical analysis on positioning precision for each method.

Performance in all Challenging Conditions. The detail positioning accuracy of different methods in challenging conditions is shown in Fig. 12. We divide the challenging conditions into two categories, namely the invisible conditions (occluded, complex background) and the confusing conditions (obscure, dense). While the current methods show certain adaptability in confusing conditions at a cost of reduced recall, they show high precision and recall decrease in the invisible conditions. Our method show higher positioning accuracy due to the ability to use the spatiotemporal sensing information and the UAV motion prediction ability.

This article has been accepted for publication in IEEE Journal on Selected Areas in Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSAC.2024.3414610

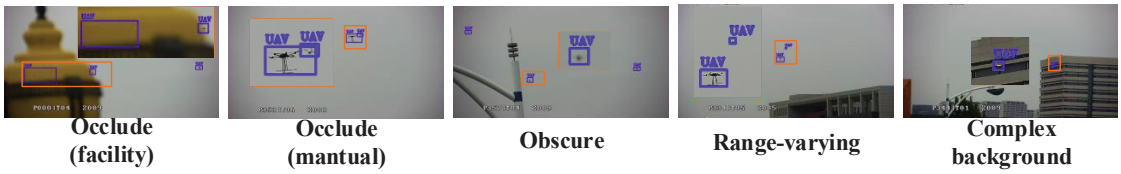

Fig. 11. The visualization results of ASTNet in challenging conditions.

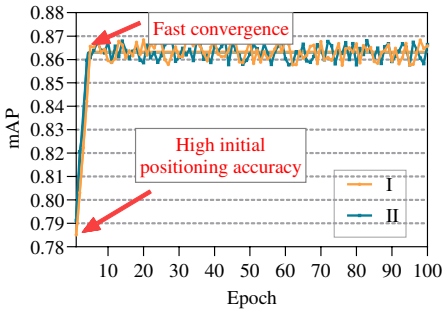

Fig. 14. The transfer training performance based on the divided groups. I: transfer from path I data to path II data. II: transfer from path II data to path I data.

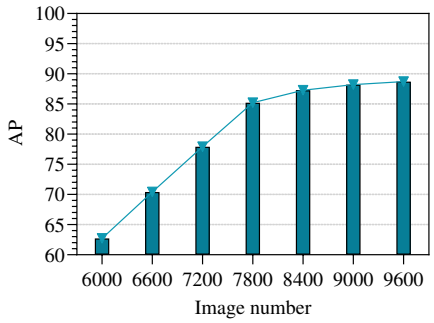

Fig. 15. The training performance w.r.t. training image number.

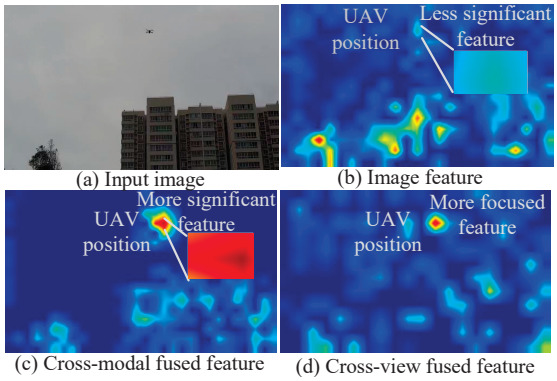

Fig. 16. Visualization of ASTNet feature fusion effectiveness.

Performance in Different Occluding Types. Since frequent occluding to avoid positioning is a key feature for clustering UAVs, we compare the positioning performance of ASTNet in different occluding types in Fig. 13, including mutual occlusion, small facility occlusion and long-term building occlusion. It shows that, due to the small size of obstacles, mutual occlusion and small facility occlusion have minor affect on ASTNet, where ASTNet obtains spatiotemporal fusion ability and can UAV patrol vehicles to provide new sensing view to overcome such occlusion. For long-term occlusion, ASTNet can predict the UAV positions in short-term (<15s) while successfully give the possible UAV flight range in the long-term (>15s), which lead to success view change of patrol vehicles. Nevertheless, the patrol vehicle can autonomous avoiding the obstacles on the sensing link by adjusting the patrol path and providing new sensing views. Therefore, our system shows strong adaptability to the frequent occlusion brought by clustering UAVs.

3) Generalization Ability.: We further provide the generalization ability of our method by dividing the collected data stream into two groups based on the patrol path for transference tests. As is shown in Fig. 14, the proposed ASTNet obtains a high positioning accuracy when transferred into a new group of data. Furthermore, the small difference between the initial and final transfer result along with the fast convergence of transfer training prove that our method has high generalization ability.

The training images requirement of ASTNet is provided in Fig. 15. We evaluate the requirement by testing the mAP trained under different training image number settings (6,000-9,600). It is shows that the training performance of ASTNet converges after the training image number is above 8,400. This proves that our method does not requires a large number of label images, and thus have a higher generalization ability for deploying in new scenarios.

4) Ablation Study: **Visualization Results on Feature Enhancement Ability.** The effectiveness of the ASTNet feature enhancement is evaluated by comparing the image feature and the enhanced feature in Fig. 16. The following observations can be found which confirms the effectiveness of our fusion method. First, Fig. 16 (c) shows higher feature saliency in the UAV area while the background significance is reduced, which illustrates that our crss-modal fusion method is adaptive and can provide precise features. Meanwhile, the cross-view fusion further shrinks the potential UAV area, which confirms the effectiveness of fusion the spatiotemporal UAV motion information.

Effectiveness of each unit of ASTNet. We further provide the ASTNet positioning accuracy with respect to each unit in Tab. VI. It is shown that the cross-modal fusion achieved by our method increases the recall by 10.6%, which not only provides rich cross-modal UAV perceptual features, but also maintains UAV physical positional information for further cross-view fusion. The cross-view fusion unit improves the mAP by 23.2%, due to the ability to provide the occluded UAV information. Finally, the 3D positioning unit provides an mAP improvement by 18.1%, since the cross-view fusion unit relies on it to provide the cross-view information. Meanwhile, we

TABLE VI
THE POSITIONING PERFORMANCE W.R.T. ASTNET UNIT.

	Precision	Recall	mAP	FPS
Proposed ASTNet	93.0%	83.7%	88.7%	20.2
–without cross-modal fusion	82.7%	73.1%	76.2%	22.6
–without cross-view fusion	76.1%	60.7%	66.5%	23.2
–without cross-domain positioning	83.8%	77.6%	80.6%	25.4

observe that our method consumes similar time on each unit, where the fusion process consumes most computation. The cross-domain positioning consumes slightly more time since traversing feature maps to acquire UAV positions requires more time. However, we can also conclude that each unit remains computational-efficient, since each of them brings a small FPS loss.

VI. DISCUSSION

Clustering UAVs v.s. Multiple UAVs. The proposed system demonstrates higher adaptability in positioning clustering UAVs compared to existing methods although they can locate multiple UAVs. This is because clustering UAVs possess unique characteristics versus multiple UAV scenarios, presenting greater positioning difficulties [52]–[55]. First, clustering UAVs exhibit superior collaborative capacities and can avoid positioning/monitoring via various formations (e.g., occluding UAVs via the leading UAV or field-of-view objects), making them more covert than non-cooperative multiple UAVs. Thus, clustering UAVs positioning requires overcoming high concealment, whereas current methods aren't specifically designed. Moreover, clustering UAVs frequently exhibit higher higher flexibility versus multiple UAVs. Namely, multiple UAVs can be long-range-distributed due to different owners. However, clustering UAVs can have either closer or greater spacing due to formation. Therefore, clustering UAVs positioning requires greater fine-grained positioning ability and and higher adaptability to better distinguish the frequency adjusting UAVs.

Patrol-Vehicle-Based System Advantages. Our patrol-vehicle-based system obtains the advantages of wide-range patrolling, high flexibility, diverse load capacity, and long endurance. It enables carrying various sensors and computing devices for long-term, wide-area, and flexible UAV positioning. Meanwhile, our system obtains finer granularity positioning for large-scale clustering UAVs through spatiotemporal fusion of dynamic perspectives based on the proposed ASTNet, while achieving real-time performance through onboard computing. In this case, our system demonstrates superior performance in frequently occluded clustering UAVs positioning. In short term, our system predicts UAVs location utilizing spatial-temporal cross-modal fusion, mitigating obstruction impact. In prolonged obstruction, our system provides UAV flight range estimate and patrol vehicles actively adjust perception position for perspective change and UAV re-capture. Nevertheless, our system can be combineed with other homogeneous/heterogeneous systems like multi-patrol-vehicle collaboration or distributed fixed monitoring facilities to achieve wider monitoring coverage and reduce patrol time.

Patrol Vehicle V.S. Other Facilities. Besides patrol vehicles, we notice that other facilities are also adopted for UAV positioning, including cellular networks and UAVs, which our proposed ASTNet is also compatible with. However, these facilities faces their limitations in long-term positioning wide-range clustering UAVs. Cellular networks provide wide coverage but have static perspectives, allowing UAVs to deceive them through specific formations. Meanwhile, their deployments require on fixed infrastructure, which is difficult and costly, especially for emergencies [56], [57]. UAV-based localization offers dynamic views but has limited payload capabilities, hindering heavy computing devices, radars and high-performance visual sensors, reducing positioning accuracy. Additionally, UAVs have limited energy and short operation duration (mostly ¡30min), making long-term positioning challenging [58]. Furthermore, sensors are mounted on UAV bodies but need omnidirectional sensing (UAVs can appear at all directions), resulting in inevitable blind spots [59]. Therefore, patrol-vehicle-based system is needed to overcome their limitations.

Network Requirement. There are various vehicle communication protocols like WiFi, millimeter-wave, and 802.11bd [60]. In our design, vehicle-mounted computing devices are used, and cross modal sensors are installed on patrol vehicles, where perception results are directly transmitted from wired for real-time positioning. We notice that our system can be linked with other devices like UAV countermeasure devices, but only needs to transmit decision-level UAV position results, which is no need for special network settings. When there are no computing devices on patrol vehicles, our method supports low latency transmission of images and radar information using various vehicle communication protocols.

VII. CONCLUSIONS

A vehicle-mounted radar-vision clustering UAVs positioning system was developed to realize precise and real-time UAV positioning. Moreover, a matching-based spatiotemporal fusion framework was established to mitigate cross-modal and cross-view spatiotemporal misalignment by adaptively exploiting the cross-modal and cross-view feature correlations. Furthermore, an attention-based spatiotemporal fusion network was designed to effectively match heterogeneous features and achieve highly accurate UAV positioning. We conducted extensive experiments in our developed clustering UAVs positioning system and demonstrated that our system is superior to the benchmark systems in terms of the positioning accuracy. Additionally, the ablation studies demonstrated the effectiveness of each model of our proposed method.

REFERENCES

[1] S.-J. Chung, A. A. Paranjape, P. Dames, S. Shen, and V. Kumar, "A survey on aerial swarm robotics," *IEEE Transactions on Robotics*, vol. 34, no. 4, pp. 837–855, 2018.

[2] C. Gudla, M. S. Rana, and A. H. Sung, "Defense techniques against cyber attacks on unmanned aerial vehicles," in *Proceedings of the international conference on embedded systems, cyber-physical systems, and applications (ESCS)*. The Steering Committee of The World Congress in Computer Science, Computer , 2018, pp. 110–116.

[3] Q. Wu, J. Xu, Y. Zeng, D. W. K. Ng, N. Al-Dhahir, R. Schober, and A. L. Swindlehurst, "A comprehensive overview on 5g-and-beyond networks with uavs: From communications to sensing and intelligence," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 10, pp. 2912–2945, 2021.

[4] I. Lachow, "The upside and downside of swarming drones," *Bulletin of the atomic scientists*, vol. 73, no. 2, pp. 96–101, 2017.

[5] M. Abdelkader, S. Güler, H. Jaleel, and J. S. Shamma, "Aerial swarms: Recent applications and challenges," *Current robotics reports*, vol. 2, pp. 309–320, 2021.

[6] S. R. Ganti and Y. Kim, "Implementation of detection and tracking mechanism for small uas," in *2016 International Conference on Unmanned Aircraft Systems (ICUAS)*, 2016, pp. 1254–1260.

[7] J. Pyrgies, "The uavs threat to airport security: Risk analysis and mitigation," *Journal of Airline and Airport Management*, vol. 9, no. 2, pp. 63–96, 2019.

[8] D. He, G. Yang, H. Li, S. Chan, Y. Cheng, and N. Guizani, "An effective countermeasure against uav swarm attack," *IEEE Network*, vol. 35, no. 1, pp. 380–385, 2021.

[9] L. Kong, Z. Liu, L. Pang, and K. Zhang, "Research on uav swarm operations," in *International Conference on Man-Machine-Environment System Engineering*. Springer, 2022, pp. 533–538.

[10] D. Qi, X.-l. Liang, Z. Li, J.-q. Zhang, P.-f. Lei, and Y.-z. Zhou, "Design and implementation of ground station software for uav swarm considering geo fence," in *Advances in Guidance, Navigation and Control: Proceedings of 2020 International Conference on Guidance, Navigation and Control, ICGNC 2020, Tianjin, China, October 23–25, 2020*. Springer, 2022, pp. 369–379.

[11] Y. Wang, T. Sun, G. Rao, and D. Li, "Formation tracking in sparse airborne networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2000–2014, 2018.

[12] S. A. H. Mohsan, N. Q. H. Othman, Y. Li, M. H. Alsharif, and M. A. Khan, "Unmanned aerial vehicles (uavs): Practical aspects, applications, open challenges, security issues, and future trends," *Intelligent Service Robotics*, vol. 16, no. 1, pp. 109–137, 2023.

[13] P. Ramesh and J. Jeyan, "Comparative analysis of the impact of operating parameters on military and civil applications of mini unmanned aerial vehicle (uav)," in *AIP conference proceedings*, vol. 2311, no. 1. AIP Publishing, 2020.

[14] W. Wu, F. Zhou, B. Wang, Q. Wu, C. Dong, and R. Q. Hu, "Unmanned aerial vehicle swarm-enabled edge computing: Potentials, promising technologies, and challenges," *IEEE Wireless Communications*, vol. 29, no. 4, pp. 78–85, 2022.

[15] M. Lehto and B. Hutchinson, "Mini-drones swarms and their potential in conflict situations," in *15th international conference on cyber warfare and security*, vol. 12, 2020, pp. 326–334.

[16] M. Y. Arafat and S. Moh, "Localization and clustering based on swarm intelligence in uav networks for emergency communications," *IEEE Internet of Things Journal*, vol. 6, no. 5, pp. 8958–8976, 2019.

[17] D. Pang, T. Shan, P. Ma, W. Li, S. Liu, and R. Tao, "A novel spatiotemporal saliency method for low-altitude slow small infrared target detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2021.

[18] Y. Dobrev, Y. Dobrev, P. Gulden, M. Lipka, T. Pavlenko, D. Moormann, and M. Vossiek, "Radar-based high-accuracy 3d localization of uavs for landing in gnss-denied environments," in *2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*. IEEE, 2018, pp. 1–4.

[19] Z. Geng, R. Xu, and H. Deng, "Lte-based multistatic passive radar system for uav detection," *IET Radar, Sonar & Navigation*, vol. 14, no. 7, pp. 1088–1097, 2020.

[20] G. Wu, F. Zhou, C. Meng, and X.-Y. Li, "Precise uav mmw-vision positioning: A modal-oriented self-tuning fusion framework," *IEEE Journal on Selected Areas in Communications*, pp. 1–1, 2023.

[21] C. Wang, T. Wang, E. Wang, E. Sun, and Z. Luo, "Flying small target detection for anti-uav based on a gaussian mixture model in a compressive sensing domain," *Sensors*, vol. 19, no. 9, p. 2168, 2019.

[22] S. Goel, "A distributed cooperative uav swarm localization system: Development and analysis," in *Proceedings of the 30th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2017)*, 2017, pp. 2501–2518.

[23] D. He, H. Liu, S. Chan, and M. Guizani, "How to govern the non-cooperative amateur drones?" *IEEE Network*, vol. 33, no. 3, pp. 184–189, 2019.

[24] J. Xiong, Z. Xiong, J. W. Cheong, J. Xu, Y. Yu, and A. G. Dempster, "Cooperative positioning for low-cost close formation flight based on relative estimation and belief propagation," *Aerospace Science and Technology*, vol. 106, p. 106068, 2020.

[25] K. Tahar and S. Kamarudin, "Uav onboard gps in positioning determination." *International Archives of the Photogrammetry, Remote Sensing & Spatial Information Sciences*, vol. 41, 2016.

[26] M. Daakir, M. Pierrot-Deseilligny, P. Bosser, F. Pichard, C. Thom, Y. Rabot, and O. Martin, "Lightweight uav with on-board photogrammetry and single-frequency gps positioning for metrology applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 127, pp. 115–126, 2017.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] A. Moffat, "Huffman coding," *ACM Computing Surveys (CSUR)*, vol. 52, no. 4, pp. 1–35, 2019.

[29] R. Opromolla, G. Fasano, and D. Accardo, "A vision-based approach to uav detection and tracking in cooperative applications," *Sensors*, vol. 18, no. 10, p. 3391, 2018.

[30] P. Mittal, R. Singh, and A. Sharma, "Deep learning-based object detection in low-altitude uav datasets: A survey," *Image and Vision computing*, vol. 104, p. 104046, 2020.

[31] C. Wang, J. Tian, J. Cao, and X. Wang, "Deep learning-based uav detection in pulse-doppler radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2021.

[32] Y. Dobrev, Y. Dobrev, P. Gulden, M. Lipka, T. Pavlenko, D. Moormann, and M. Vossiek, "Radar-based high-accuracy 3d localization of uavs for landing in gnss-denied environments," in *2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*. IEEE, 2018, pp. 1–4.

[33] H. Lee, W. Kim, and J. Seo, "Simulation of uwb radar-based positioning performance for a uav in an urban area," in *2018 IEEE International Conference on Consumer Electronics-Asia (ICCE-Asia)*. IEEE, 2018, pp. 206–212.

[34] W. Chen, J. Liu, and J. Li, "Classification of uav and bird target in low-altitude airspace with surveillance radar data," *The Aeronautical Journal*, vol. 123, no. 1260, pp. 191–211, 2019.

[35] S. Yuan, Y. Yang, T. H. Nguyen, T.-M. Nguyen, J. Yang, F. Liu, J. Li, H. Wang, and L. Xie, "Mmaud: A comprehensive multi-modal anti-uav dataset for modern miniature drone threats," 2024.

[36] H. Cheng, L. Lin, Z. Zheng, Y. Guan, and Z. Liu, "An autonomous vision-based target tracking system for rotorcraft unmanned aerial vehicles," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 1732–1738.

[37] Y. Zheng, Z. Chen, D. Lv, Z. Li, Z. Lan, and S. Zhao, "Air-to-air visual detection of micro-uavs: An experimental evaluation of deep learning," *IEEE Robotics and automation letters*, vol. 6, no. 2, pp. 1020–1027, 2021.

[38] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[39] J. Hosang, R. Benenson, and B. Schiele, "Learning Non-Maximum Suppression," in *2017 IEEE Conf. Comput. Vision and Pattern Recognition (CVPR)*, 2017, pp. 4507–4515.

[40] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "Fastdepth: Fast monocular depth estimation on embedded systems," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6101–6108.

[41] D. R. I. M. Setiadi, "Psnr vs ssim: imperceptibility quality assessment for image steganography," *Multimedia Tools and Applications*, vol. 80, no. 6, pp. 8423–8444, 2021.

[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[43] "Yolov5-lite," [online] Available: https://github.com/ppogg/YOLOv5-Lite.

[44] W. Zhan, C. Sun, M. Wang, J. She, Y. Zhang, Z. Zhang, and Y. Sun, "An improved yolov5 real-time detection method for small objects captured by uav," *Soft Computing*, vol. 26, pp. 361–373, 2022.

[45] X. Farhodov, O.-H. Kwon, K. W. Kang, S.-H. Lee, and K.-R. Kwon, "Faster rcnn detection based opencv csrt tracker using drone data," in *2019 International Conference on Information Science and Communications Technologies (ICISCT)*. IEEE, 2019, pp. 1–3.

[46] T. Liu, Z. Yang, A. Marino, G. Gao, and J. Yang, "Robust cfar detector based on truncated statistics for polarimetric synthetic aperture radar," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 9, pp. 6731–6747, 2020.

[47] D. Kumuda, G. Vandana, B. Pardhasaradhi, B. Raghavendra, P. Srihari, and L. R. Cenkeramaddi, "Multi target detection and tracking by mitigating spot jammer attack in 77ghz mm-wave radars: An experimental evaluation," *IEEE Sensors Journal*, 2022.

[48] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE, 2019, pp. 1–7.

[49] T.-Y. Lim, A. Ansari, B. Major, D. Fontijne, M. Hamilton, R. Gowaikar, and S. Subramanian, "Radar and camera early fusion for vehicle detection in advanced driver assistance systems," in *Machine Learning for Autonomous Driving Workshop at the 33rd Conference on Neural Information Processing Systems*, vol. 2, 2019, p. 7.

[50] M. Dreher, E. Erçelik, T. Bänziger, and A. Knoll, "Radar-based 2d car detection using deep neural networks," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–8.

[51] R. Padilla, S. L. Netto, and E. A. B. da Silva, "A Survey on Performance Metrics for Object-Detection Algorithms," in *2020 Int. Conf. Syst., Signals and Image Processing (IWSSIP)*, 2020, pp. 237–242.

[52] J. Yan, H. Xie, and J. Li, "Modeling and optimization of deploying anti-uav swarm detection systems based on the mixed genetic and monte carlo algorithm," in *2021 IEEE International Conference on Unmanned Systems (ICUS)*. IEEE, 2021, pp. 773–779.

[53] Q. Hao, W. Li, Z. Qiu, and J. Zhang, "Research on anti uav swarm system in prevention of the important place," in *Journal of Physics: Conference Series*, vol. 1507, no. 5. IOP Publishing, 2020, p. 052020.

[54] W. Yi, Y. Liu, Y. Deng, and A. Nallanathan, "Clustered uav networks with millimeter wave communications: A stochastic geometry view," *IEEE Transactions on Communications*, vol. 68, no. 7, pp. 4342–4357, 2020.

[55] H. Zhao, H. Wang, W. Wu, and J. Wei, "Deployment algorithms for uav airborne networks toward on-demand coverage," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 2015–2031, 2018.

[56] U. Challita, H. Ryden, and H. Tullberg, "When machine learning meets wireless cellular networks: Deployment, challenges, and applications," *IEEE Communications Magazine*, vol. 58, no. 6, pp. 12–18, 2020.

[57] R. Borralho, A. Mohamed, A. U. Quddus, P. Vieira, and R. Tafazolli, "A survey on coverage enhancement in cellular networks: Challenges and solutions for future deployments," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1302–1341, 2021.

[58] S. A. H. Mohsan, N. Q. H. Othman, Y. Li, M. H. Alsharif, and M. A. Khan, "Unmanned aerial vehicles (uavs): Practical aspects, applications, open challenges, security issues, and future trends," *Intelligent Service Robotics*, vol. 16, no. 1, pp. 109–137, 2023.

[59] K. Messaoudi, O. S. Oubbati, A. Rachedi, A. Lakas, T. Bendouma, and N. Chaib, "A survey of uav-based data collection: Challenges, solutions and future perspectives," *Journal of Network and Computer Applications*, p. 103670, 2023.

[60] S. Zeadally, M. A. Javed, and E. B. Hamida, "Vehicular communications for its: Standardization and challenges," *IEEE Communications Standards Magazine*, vol. 4, no. 1, pp. 11–17, 2020.

Guangyu Wu (Member, IEEE) is currently pursuing the M.S. degree with the Department of Computer Science and Technology, University of Science and Technology of China. He received the B.S. degree with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics. He received the National Scholarship in 2022 and 2023. His research interests include machine learning, mobile computing and the Internet of Things.

Fuhui Zhou (Senior Member, IEEE) is currently a Full Professor at Nanjing University of Aeronautics and Astronautics. He is an IEEE Senior Member. His research interests focus on cognitive radio, RF machine learning, knowledge graph, edge computing, and resource allocation. He was awarded as Young Elite Scientist Award of China and URSI GASS Young Scientist Award. He serves as an Editor of IEEE Transactions on Communications, IEEE Systems Journal, IEEE Wireless Communications Letters, IEEE Access and Physical Communications.

Kai Kit Wong (Fellow, IEEE) received the BEng, the MPhil, and the PhD degrees, all in Electrical and Electronic Engineering, from the Hong Kong University of Science and Technology, Hong Kong, in 1996, 1998, and 2001, respectively. After graduation, he took up academic and research positions at the University of Hong Kong, Lucent Technologies, Bell-Labs, Holmdel, the Smart Antennas Research Group of Stanford University, and the University of Hull, UK. He is Chair in Wireless Communications at the Department of Electronic and Electrical Engineering, University College London, UK. His current research centers around 6G and beyond mobile communications. He is Fellow of IEEE and IET. He served as the Editor-in-Chief for IEEE Wireless Communications Letters between 2020 and 2023.

Xiang-Yang Li (Fellow, IEEE) is a professor and Executive Dean at School of Computer Science and Technology, USTC and co-Chair of ACM China Council. He is an ACM Fellow (2019), IEEE fellow (2015), an ACM Distinguished Scientist (2014). He was a full professor at Computer Science Department of IIT. Dr. Li received M.S. (2000) and Ph.D. (2001) degree at Department of Computer Science from University of Illinois at Urbana-Champaign. He received a Bachelor degree at Department of Computer Science from Tsinghua University, P.R. China, in 1995. His research interests include Artificial Intelligence of Things(AIOT), privacy and security of AIOT, and data sharing and trading.