Chinese Journal of Aeronautics, (2024), 37(6): 219–231

Chinese Society of Aeronautics and Astronautics
& Beihang University

Chinese Journal of Aeronautics

cja@buaa.edu.cn
www.sciencedirect.com

FULL LENGTH ARTICLE

Absolute pose estimation of UAV based on large-scale satellite image

Hanyu WANG [a,b], **Qiang SHEN** [a,b,*], **Zilong DENG** [a], **Xinyi CAO** [a], **Xiaokang Wang** [a]

[a] *School of Mechatronical Engineering, Beijing Institute of Technology, Beijing 100081, China*
[b] *Beijing Institute of Technology Chongqing Innovation Center, Chongqing 401120, China*

Received 23 May 2023; revised 20 June 2023; accepted 8 September 2023
Available online 26 December 2023

KEYWORDS

Satellite imagery;
Unmanned Aerial Vehicle (UAV);
Deep neural networks;
Vision navigation;
von Mises-Fisher distribution

Abstract Obtaining absolute pose based on pre-loaded satellite images is one of the important means of autonomous navigation for small Unmanned Aerial Vehicles (UAVs) in Global Navigation Satellite System (GNSS) denied environments. Most of the previous works have tended to build Convolutional Neural Networks (CNNs) to extract features and then directly regress the pose, which will fail when solving the challenges caused by the huge viewpoint and size differences between "UAV-satellite" image pairs in real-world scenarios. Therefore, this paper proposes a probability distribution/regression integrated deep model with the attention-guided triple fusion mechanism, which estimates discrete distributions in pose space and three-dimensional vectors in translation space. In order to overcome the shortage of the relevant dataset, this paper simulates image datasets captured by UAVs with forward-facing cameras during target searching and autonomous attacking. The effectiveness, superiority, and robustness of the proposed method are verified by simulated datasets and flight tests.

1. Introduction

In recent years, outdoor UAVs have seen exponential growth in commercial, [1] military, [2] and civil [3] applications, such as search and rescue, [4] reconnaissance, [5] and suicide attacks. [6] Almost all of these applications require accurate absolute pose estimation of the UAV during flight, which refers to finding the position of the UAV in the world frame and the attitude in the navigation frame. Nowadays, absolute pose estimation mainly relies on satellite/inertial integrated navigation systems in outdoor large-scale scenarios. [7–8] However, satellite signals are naturally susceptible to multi-path interference, intentional jamming, or spoofing by an adversary; the low-cost inertial navigation errors increase sharply over time without the assistance of GNSS and will cause catastrophic ramifications. In such cases, with the development of image processing and computer vision, lightweight, low-cost, high-accuracy, and

* Corresponding author.
E-mail address: bit82shen@bit.edu.cn (Q. SHEN).

Peer review under responsibility of Editorial Committee of CJA.

Production and hosting by Elsevier

https://doi.org/10.1016/j.cja.2023.12.028

fast-computing vision sensors are desired to aid navigation to improve the survivability of UAVs.

Although visual odometry has been extensively studied and formed mature theory, [9] it has not yet reached a practical level, especially in outdoor long trajectory conditions. Worse still, it cannot provide the absolute pose. With the ample, free, GPS-aligned, high-resolution, open-source satellite images covering the vast majority of the world, [10] the method of UAV absolute pose estimation based on pre-loaded satellite images has become a hotspot. The main advantage of the vision-based absolute pose estimation is its inherent immunity to drift over time. However, images from UAVs and satellites are often taken by different sensors, at different time, with different levels of illumination and different resolutions. Estimating relative pose from heterologous and significantly different image pairs is the core of vision-based absolute pose estimation, which makes the methods based on a template or sparse feature descriptors often fail. [11] Thus, researchers have focused on deep learning-based methods for feature match and relative pose estimation recently. Bianchi et al. [12] first pre-encoded the satellite images on the desired flight path based on the deep learning model and then encoded the UAV images during the flight in the same model, to obtain the pose through the comparison with the satellite image features. Wu et al. [13] proposed a multi-stage Lucas-Kanade deep learning algorithm to refine the UAV geo-localization results iteratively, making full use of the global texture information. Sui et al. [14] designed a fast detector-free two-stage matching method based on a perspective transformation module and the minimum Euclidean distance, which acquired great improvement in the robustness of the weak texture and viewpoint data. However, these methods depend on the assumption of a downward-facing camera, which is a major limitation: Either installing an additional camera, which imposes a burden on cost, computational power, and weight; or equipping the gimbal to adjust the viewing angle at intervals, which causes a trade-off between the autonomous navigation and primary tasks, such as safe flight and target searching. [15].

There are two main solutions to the problem of the cross-view pose estimation from "UAV-satellite" image pairs. Firstly, the camera image is ortho-projected based on the principle of perspective projection transformation and CNN focuses on learning the feature-matching relationship between the aligned image pairs. Kinnari et al. [16] proposed a matching metric between the orthorectified UAV and satellite images, which relaxes the requirement of the downward-facing camera, but strictly relies on the assumption of local planarity of the environment, thus not suitable for areas with large terrain and building undulations, such as urban, mountainous, and suburban. Tian et al. [17] introduced a generative adversarial network to correct the UAV images and then trained a CNN to estimate pose, which imposes higher requirements on the precision acquisition of initial pose. Nowadays, researchers preferred the second end-to-end learning approach, [18–19] restricted by the resolution of the onboard camera and computational cost. This method usually employs a Siamese-CNN framework [20] to extract view-invariant features. Shetty et al. [21] built a cross-view geolocation network to acquire the weighted absolute pose estimation of UAV, consisting of the scene localization network and camera localization network. Ding et al. [22] proposed a cross-view matching method based on building classification. Chen et al. [23] pre-encoded the terrain map of the known flight area and used a combination of CNN and Perspective-*n*-Point solvers to obtain a refined UAV pose during flight. However, these methods only consider a small viewpoint difference, where the camera pitch angle $\in [-90^{\circ}, -45^{\circ}]$. The challenging scenario where the camera is fully forward-facing is not involved. Meanwhile, the pre-loaded satellite images are all cropped to a small size to acquire accurate matching, resulting in the large storge. Therefore, the absolute pose estimation method based on pre-loaded large-scale satellite images for UAV is still in the preliminary exploration stage.

Previous works [24–26] show that estimating a discrete distribution, or a heatmap over a quantized output space, is consistently superior to direct regression on the continuous space. Inspired by this, researchers proposed probabilistic depth models to estimate the relative poses between image pairs. Chen et al. [27] decomposed the relative pose into direction vectors and then predicted the discrete distribution, improving the estimation accuracy. However, the method needs to decouple the image information caused by the translation and attitude change, and the predicted translation is scale-free, which makes the method complex and impractical.

In order to solve the difficulty of global pose estimation based on large-scale satellite images for small UAVs equipped with forward-facing cameras, this paper proposes a probability distribution/regression integrated deep model with the attention-guided triple fusion mechanism, namely, Pd/Re-Net with AGTFM. The proposed model enables autonomous global navigation of small UAVs in GNSS-denied environments. The main contributions of this paper are presented as follows:

(1) An absolute pose estimation method of UAV based on large-scale satellite images is proposed, which is a deep learning model combining probability distribution branch and direct regression branch. By analyzing the continuity of different attitude characterization parameters during UAV search and attack targets, an attitude estimation branch based on von Mises-Fisher (vMF) distribution has been established. Considering the balance between the improved performance and increased model complexity, a translation estimation branch based on fully connected layers has been established.
(2) An attention-guided triple fusion mechanism is proposed to improve the feature extraction module. By assigning higher weights to the viewpoint-invariant feature regions and introducing the local detail features, the process of "focus and zoom in" of human eyes is simulated to enhance the estimation accuracy.
(3) Considering that the real-scene UAV dataset is very scarce, we simulate the flight data of UAVs with various poses using Google Earth Studio and quantitatively prove the performance of the proposed method. Furthermore, the flight test has been conducted to verify the validity and robustness.

The rest of the paper is structured as follows. Section 2 illuminates the structure of the proposed Pd/Re-Net with AGTFM model in detail, including the parameterization of the discrete distribution, the attention-guided triple fusion mechanism, and the translation estimation module. Section 3

introduces the dataset built in this paper, as well as the results of the simulation and the flight experiment. Section 4 concludes the paper.

2. Method

2.1. Overview

Given a set of pre-loaded large-scale satellite images of the target area and the images captured by the UAV during flight, the goal is to directly obtain the relative pose between the image pairs through an end-to-end deep learning model, and then acquire the absolute pose of the UAV. The framework of the UAV absolute pose estimation method proposed in this paper is shown in Fig. 1. Firstly, the Siamese architecture is constructed, where two feature extraction networks have the same structure and share the same weights to process the input images. Inspired by PoseNet, [28] the feature extraction network is based on the GoogleNet architecture [29], with 22 CNN layers and 6 inception modules. In our Pd/Re-Net with AGTFM model, the last classification layer of GoogleNet is changed to the fully connected layers, and the attention-guided triple fusion mechanism is introduced to improve the estimation accuracy. Next, the extracted high-dimensional image embeddings are concatenated and input into two different branches of attitude estimator and translation estimator respectively. The attitude estimator branch is mainly composed of a spherical decoder [27], which maps the image pairs to the discrete probability distribution of the column vectors in the rotation matrix. The expectations of the distribution are regarded as the prediction of the rotation matrix. The translation estimator branch is composed of two fully connected layers, and the translation vector is obtained by direct regression.

2.2. Attitude estimator

2.2.1. Attitude representation parameters

Attitude transformation can be represented with quaternions, Euler angles, rotation matrix, and equivalent rotation vector, whose continuity will have an important impact on the fitting ability of the neural network. Therefore, we analyze the continuity variation of these representations when the UAV performs a combat task near the target. At this time, the roll γ changes slightly, the pitch $\theta \in [-50^{\circ}, 0^{\circ}]$ and the head $\psi \in [-180^{\circ}, 180^{\circ}]$.

First, the discontinuity of Euler angle is very intuitive and easy to find when $\psi = -180^{\circ}$ or $\psi = 180^{\circ}$.

Second, when the East-North-Up (ENU) frame is taken as the navigation frame (n), the rotation matrix between the body frame (b) and the navigation frame can be defined as

$$\boldsymbol{R}_b^n = \begin{bmatrix} \mathrm{c}_\psi \mathrm{c}_\gamma - \mathrm{s}_\psi \mathrm{s}_\theta \mathrm{s}_\gamma & -\mathrm{s}_\psi \mathrm{c}_\theta & \mathrm{c}_\psi \mathrm{s}_\gamma + \mathrm{s}_\psi \mathrm{s}_\theta \mathrm{c}_\gamma \\ \mathrm{s}_\psi \mathrm{c}_\gamma + \mathrm{c}_\psi \mathrm{s}_\theta \mathrm{s}_\gamma & \mathrm{c}_\psi \mathrm{c}_\theta & \mathrm{s}_\psi \mathrm{s}_\gamma - \mathrm{c}_\psi \mathrm{s}_\theta \mathrm{c}_\gamma \\ -\mathrm{c}_\theta \mathrm{s}_\gamma & \mathrm{s}_\theta & \mathrm{c}_\theta \mathrm{c}_\gamma \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} & R_{13} \\ R_{21} & R_{22} & R_{23} \\ R_{31} & R_{32} & R_{33} \end{bmatrix} \tag{1}$$

where $\mathrm{s}_X = \sin X, \mathrm{c}_X = \cos X, (X = \theta, \gamma, \psi)$.Then, the rotation matrix between UAV and satellite images is expressed as the product of multiple continuous representations:

$$\boldsymbol{R}_{\mathrm{uav}}^{\mathrm{sat}} = \boldsymbol{R}_{\mathrm{uav}}^{n} \times \left(\boldsymbol{R}_{\mathrm{sat}}^{n}\right)^{\mathrm{T}} \tag{2}$$

which is continuous undoubtedly.

Next, quaternions $\boldsymbol{Q}_b^n = [\,q_0 \quad q_1 \quad q_2 \quad q_3\,]^{\mathrm{T}}$ can be computed from the rotation matrix:

$$\boldsymbol{Q}_b^n = \begin{bmatrix} q_0 \\ q_1 \\ q_2 \\ q_3 \end{bmatrix} = \begin{cases} \left[\frac{R_{32}-R_{23}}{4q_1} \quad \frac{\sqrt{1+R_{11}-R_{22}-R_{33}}}{2} \quad \frac{R_{12}+R_{21}}{4q_1} \quad \frac{R_{13}+R_{31}}{4q_1}\right]^{\mathrm{T}}, R_{11} \geqslant R_{22}+R_{33} \\ \left[\frac{R_{13}-R_{31}}{4q_2} \quad \frac{R_{12}+R_{21}}{4q_2} \quad \frac{\sqrt{1-R_{11}+R_{22}-R_{33}}}{2} \quad \frac{R_{23}+R_{32}}{4q_2}\right]^{\mathrm{T}}, R_{22} \geqslant R_{11}+R_{33} \\ \left[\frac{R_{21}-R_{12}}{4q_3} \quad \frac{R_{13}+R_{31}}{4q_3} \quad \frac{R_{23}+R_{32}}{4q_3} \quad \frac{\sqrt{1-R_{11}-R_{22}+R_{33}}}{2}\right]^{\mathrm{T}}, R_{33} \geqslant R_{11}+R_{22} \\ \left[\frac{\sqrt{1+R_{11}+R_{22}+R_{33}}}{2} \quad \frac{R_{32}-R_{23}}{4q_0} \quad \frac{R_{13}-R_{31}}{4q_0} \quad \frac{R_{21}-R_{12}}{4q_0}\right]^{\mathrm{T}}, \text{others} \end{cases} \tag{3}$$

$\boldsymbol{Q}_{\mathrm{uav}}^{\mathrm{sat}}$ is given as the product of two quaternions:

$$\boldsymbol{Q}_{\mathrm{uav}}^{\mathrm{sat}} = \boldsymbol{Q}_{\mathrm{uav}}^{n} \otimes \bar{\boldsymbol{Q}}_{\mathrm{sat}}^{n} \tag{4}$$

where the quaternion from the satellite image to navigation frame $\boldsymbol{Q}_{\mathrm{sat}}^{n}$ is constant:

$$\boldsymbol{Q}_{\mathrm{sat}}^{n} = [\,q_0 \quad q_1 \quad q_2 \quad q_3\,]^{\mathrm{T}} = [\,-0.7071 \quad 0.7071 \quad 0 \quad 0\,]^{\mathrm{T}} \tag{5}$$

As described in Eq. (3), a discontinuity point occurs when the judgment condition changes. During the flight of the UAV for the search and attack mission, the discontinuity always occurs when $R_{33} \geqslant R_{11} + R_{22}$ is no longer satisfied, that is,

$$\mathrm{c}_\psi > \frac{\mathrm{c}_\theta \mathrm{c}_\gamma (\mathrm{c}_\theta + \mathrm{c}_\gamma) - \sqrt{(1 + 2\mathrm{c}_\theta \mathrm{c}_\gamma)\mathrm{s}_\theta^2 \mathrm{s}_\gamma^2}}{(1 + \mathrm{c}_\theta \mathrm{c}_\gamma)^2} \tag{6}$$

An illustration is shown in Fig. 2, where the quaternion changes with the head angle when $\theta = -20^{\circ}, \gamma = -10^{\circ}$. From Eq. (6), the intermittent point occurs when $\psi \approx -63.03^{\circ}$, which is consistent with the results shown in Fig. 2.

The equivalent rotation vector can be obtained through quaternions:

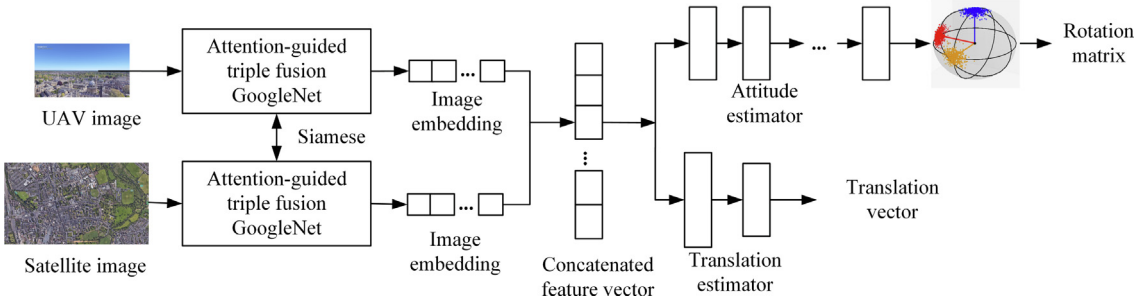


Fig. 1 Overview of proposed Pd/Re-Net with AGTFM model.

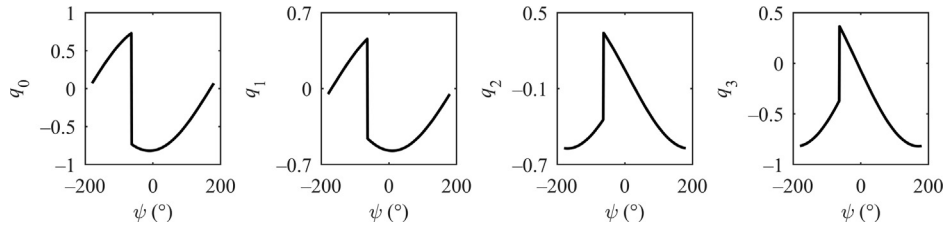


Fig. 2 Quaternion variation with yaw angle.

$$(\boldsymbol{r}_{ev})_b^n = \frac{2\arccos(q_0)}{\sin(\arccos(q_0))}\begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix} \tag{7}$$

where it is also discontinuous.

Consequently, the rotation matrix is chosen to represent the attitude, of which the probability distribution and discretization will be further analyzed.

2.2.2. von Mises-Fisher distribution

The rotation matrix $\boldsymbol{R} \in \mathrm{SO}(\mathbf{3})$ can be split into three column vectors $\boldsymbol{R} = [\,\boldsymbol{r}_1 \quad \boldsymbol{r}_2 \quad \boldsymbol{r}_3\,]$, where each component is a unit vector. Thus, the task of relative attitude prediction between "UAV-satellite" image pairs can be converted to estimate a set of direction vectors $\boldsymbol{r}_1, \boldsymbol{r}_2, \boldsymbol{r}_3 \in S^2$. Assuming that the vectors are independent of each other, we can predict the probability distribution of each vector over the S^2 space and then use the expectation of the distribution as the predicted value of the vectors. The von Mises-Fisher distribution is a continuous probability distribution on the $(n-1)$ dimensional sphere S^{n-1}, which can be seen as a spherical simulation of a normal distribution. [30] The density of the von Mises-Fisher distribution for a n-dimensional random unit vector, $\boldsymbol{x}$, is given by

$$p(\boldsymbol{x}, \boldsymbol{\mu}, \kappa) = \frac{\kappa^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)} \cdot \mathrm{e}^{\kappa \boldsymbol{\mu}^{\mathrm{T}} \boldsymbol{x}} \tag{8}$$

where $\boldsymbol{\mu}$ is the mean direction, which plays a role similar to that of the expectation in a normal distribution; κ is the concentration parameter and is non-negative, whose reciprocal plays a role similar to that of the variance in a normal distribution; [31–32] $I_{n/2-1}$ denotes the first kind of $(n/2-1)$- order modified Bessel function.

Given N samples from a vMF distribution, the mean direction can be estimated as follows:

$$\hat{\boldsymbol{\mu}} = \frac{\sum_{i=1}^{N} \boldsymbol{x}_i}{\| \sum_{i=1}^{N} \boldsymbol{x}_i \|_2} \tag{9}$$

In this paper, the learning problem based on the von Mises-Fisher distribution is defined as follows. Given a set of "UAV-satellite" image pairs, the objective is to learn a vMF distribution for every pair parameterized by $\{\kappa_i, \boldsymbol{\mu}_i\}(i = 1, 2, \ldots, N)$ in S^2 space. An illustration is given in Fig. 3. The top and the second rows present "UAV-satellite" image pairs with different relative attitudes, of which the 3D directional vectors from the vMF distribution on the S^2 sphere are shown in the last three rows for different values of κ. It can be observed that the concentration of the distribution around the mean direction becomes higher as κ increases.

For a set of image pairs, the ground truth rotation matrix is denoted by $\boldsymbol{R}^* = [\,\boldsymbol{r}_1^* \quad \boldsymbol{r}_2^* \quad \boldsymbol{r}_3^*\,]$. Taking $\boldsymbol{r}_1^*$ as an example, the vMF distribution with $\boldsymbol{r}_1^*$ as the mean can be calculated by Eq. (8). Then, the Fibonacci lattice method [33] is used to sample 64×64 discrete points uniformly on the sphere, and the set of the corresponding vMF distribution values of each point can be expressed as $p(\boldsymbol{r}_1^*)$. The same operations are processed for $\boldsymbol{r}_2^*$ and $\boldsymbol{r}_3^*$, acquiring $p(\boldsymbol{r}_2^*)$ and $p(\boldsymbol{r}_3^*)$. Therefore, the rotation matrix $\boldsymbol{R}^*$ can be mapped into a discrete vMF distribution over the attitude space SO(3):

$$\Gamma^*(\boldsymbol{R}^*) = \Gamma^*(\boldsymbol{r}_1^*, \boldsymbol{r}_2^*, \boldsymbol{r}_3^*) = \left\{p(\boldsymbol{r}_1^*), p(\boldsymbol{r}_2^*), p(\boldsymbol{r}_3^*)\right\} \tag{10}$$

For each direction vector in the rotation matrix, $\hat{\boldsymbol{r}}_j (j = 1, 2, 3)$, the proposed network model directly predicts the vMF distribution values of the same 64×64 points sampled on the S^2 sphere, $p(\hat{\boldsymbol{r}}_j)(j = 1, 2, 3)$. The prediction of the vMF distribution of the rotation matrix can be denoted as

$$\hat{\Gamma}(\hat{\boldsymbol{R}}) = \hat{\Gamma}(\hat{\boldsymbol{r}}_1, \hat{\boldsymbol{r}}_2, \hat{\boldsymbol{r}}_3) = \{p(\hat{\boldsymbol{r}}_1), p(\hat{\boldsymbol{r}}_2), p(\hat{\boldsymbol{r}}_3)\} \tag{11}$$

And $\Gamma^*(\boldsymbol{R}^*)$ is abbreviated as Γ^*, and $\hat{\Gamma}(\hat{\boldsymbol{R}})$ is abbreviated as $\hat{\Gamma}$ for convenience.

The estimation of the rotation matrix can be regarded as the mean value of the distribution, which can be derived from Eq. (9) as

$$\hat{\boldsymbol{R}} = [\,\hat{\boldsymbol{r}}_1 \quad \hat{\boldsymbol{r}}_2 \quad \hat{\boldsymbol{r}}_3\,] \tag{12}$$

In order to supervise both the rotation matrix and its probability density distribution simultaneously, the loss function of the proposed network is composed of two parts. One is the loss of the predicted direction vectors, which is expressed as the cosine similarity between two 3D unit vectors:

$$L_{dir}\left(\hat{\boldsymbol{R}}, \boldsymbol{R}^*\right) = \sum_{i=1}^{3}\left(1 - \frac{\left(\boldsymbol{r}_i^*\right)^{\mathrm{T}} \hat{\boldsymbol{r}}_i}{\| \boldsymbol{r}_i^* \| \| \hat{\boldsymbol{r}}_i \|}\right) \tag{13}$$

The other is the loss of the predicted distributions, which provides dense supervision for each discrete point:

$$\begin{aligned} L_{\mathrm{dis}}\left(\hat{\Gamma}, \Gamma^*\right) &= L_{\mathrm{dis}}\left(\hat{\Gamma}(\hat{\boldsymbol{R}}), \Gamma^*(\boldsymbol{R}^*)\right) \\ &= \sum_{i=1}^{3}\left(\| p(\hat{\boldsymbol{r}}_i) - p(\boldsymbol{r}_i^*) \|_2\right)^2 \end{aligned} \tag{14}$$

Hence, the total vMF loss function for rotation matrix estimation of the deep learning model is

$$L_{\mathrm{vMF}}(\hat{\boldsymbol{R}}, \hat{\Gamma}) = L_{\mathrm{dir}}\left(\hat{\boldsymbol{R}}, \boldsymbol{R}^*\right) + \alpha L_{\mathrm{dis}}\left(\hat{\Gamma}, \Gamma^*\right) \tag{15}$$

where α is the distribution loss weight.

In general, the rotation matrix predicted by the network model often fails to meet the orthogonalization requirement, so we additionally introduced the SVD algorithm for orthogonalization, [34] which has been widely used in rotation estimation tasks.

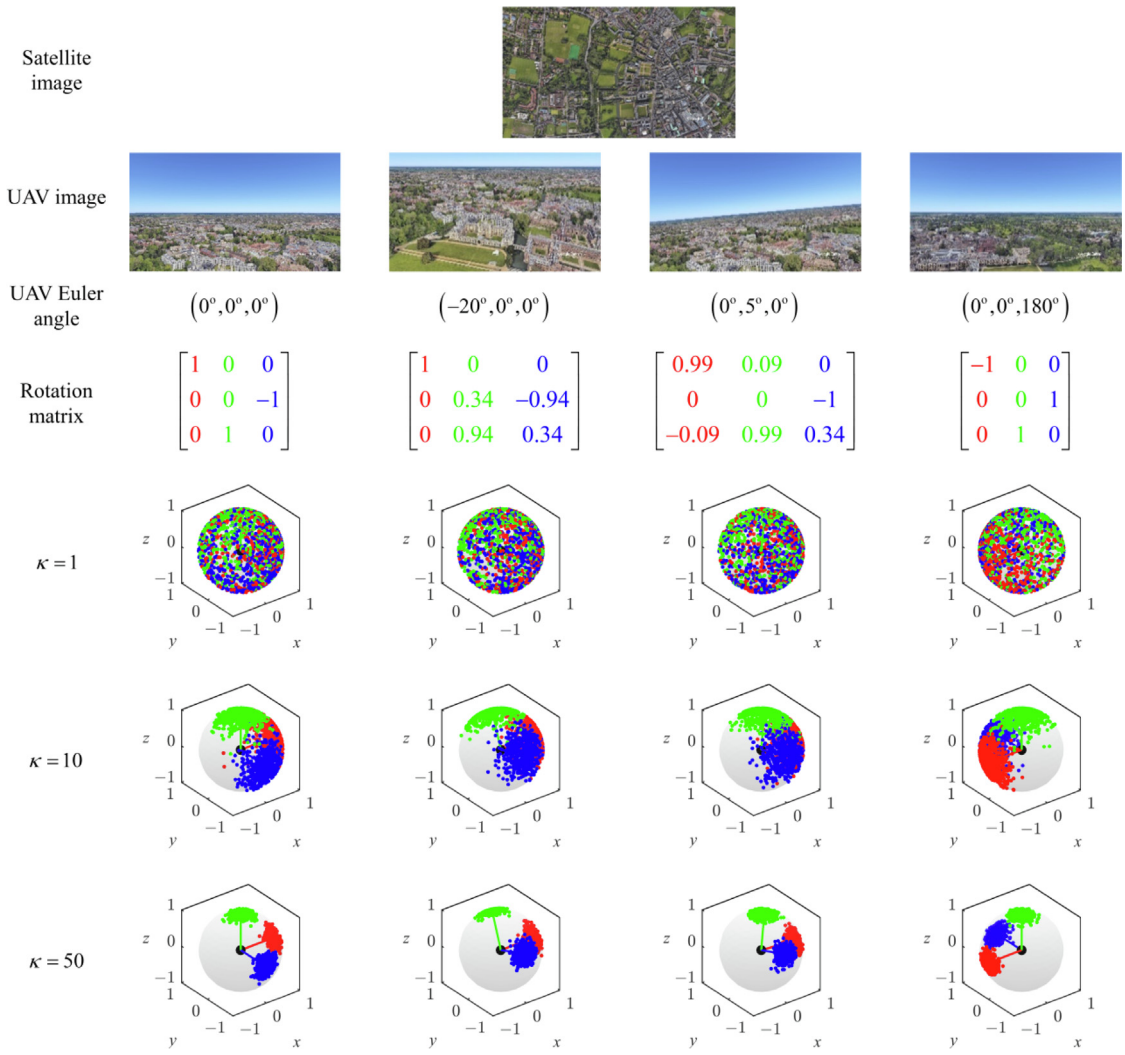


Fig. 3 Relative attitude samples and corresponding vMF distributions with different values of κ.

2.2.3. Attention-guided triple fusion mechanism

There is ample information in large-scale satellite images, and the traditional feature extraction network, GoogleNet, is difficult to evaluate the validity among various information. Thus, it is difficult to accurately extract the most relevant areas with UAV images. Worse more, after successive convolution and pooling operations, the pixel-level local features in the image embeddings extracted from GoogleNet are gradually lost. However, for the "UAV-satellite" image pairs with strong viewpoint differences, which share little information, this local feature will play an important role in estimating the relative pose.

The human eye's idea of image matching is worth learning: find the region with the highest degree of match in an image pair first, and then focus on that region and concentrate on the details. Enlightened by this thought, we propose an attention-guided triple fusion mechanism to improve the feature extraction module, achieving the process of "focus and concentrate" between image pairs to acquire more accurate pose estimation.

The optimized feature extraction network structure is shown in Fig. 4. First, in order to adaptively focus on areas generating a significant impact on the pose estimation results, the channel attention mechanism, Senet, [35] is introduced in both the two auxiliary regression modules and the final regression module in GoogleNet. Senet models the interdependencies among channels, acquires the importance of each feature channel by learning, then enhances useful features and suppresses useless features accordingly. Next, the obtained image embeddings 1–3 at different levels are all input into the attitude estimator. And the relative attitude estimation parameters with comprehensive characterizations are acquired, which include two auxiliary attitude distribution estimations and the final attitude distribution prediction of the network.

The auxiliary attitude distribution estimations are only used in the training process, optimizing the network parameters by gradient descent, so the network can pay attention to both global semantic information and local pixel details.

2.3. Loss calculation

As shown in Fig. 1, for a pair of "UAV-satellite" images, the proposed Pd/Re-Net with AGTFM model extracts feature first and then outputs two auxiliary attitude distribution

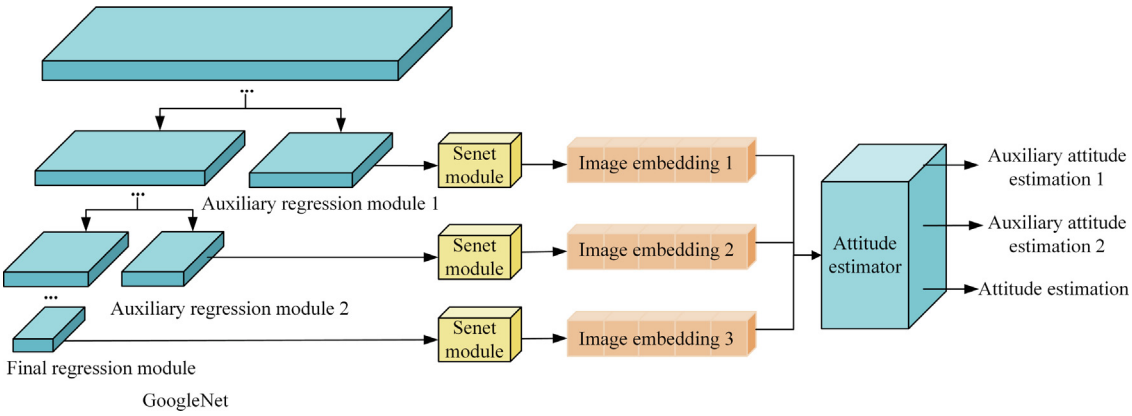


Fig. 4 Structure of proposed attention-guided triple fusion mechanism.

estimations $\hat{\Gamma}_{a,i}(\hat{r}_1,\hat{r}_2,\hat{r}_3)(i=1,2)$, final attitude distribution prediction of the network $\hat{\Gamma}_f(\hat{r}_1,\hat{r}_2,\hat{r}_3)$ and the relative translation prediction $\hat{\boldsymbol{T}}=\begin{bmatrix}\Delta\hat{E} & \Delta\hat{N} & \Delta\hat{U}\end{bmatrix}$ in the training phase, where $\Delta\hat{E},\Delta\hat{N},\Delta\hat{U}$ represents the projection of the relative translation between the drone and satellite image capture points in three directions on the navigation frame. The label of the model is the camera transformation between "satellite-UAV" image pairs, including the relative translation $\boldsymbol{T}^*=\begin{bmatrix}\Delta E^* & \Delta N^* & \Delta U^*\end{bmatrix}$ and the rotation matrix $\boldsymbol{R}^*$.

The total loss L_{total} consists of translation estimation loss L_{tran} and attitude estimation loss L_{att}. The least square error is chosen as the translation estimation loss while the attitude estimation loss includes three-level rotation matrix losses. We summarize the total loss algorithm in the training phase in Algorithm 1.

Algorithm 1. Total loss.

Input: a pair of "UAV-satellite" images, labels $(\boldsymbol{T}^*,\boldsymbol{R}^*)$
Output: L_{total}
1. Substitute $\boldsymbol{R}^*$ into Eq. (8) to obtain Γ^*, where the column vectors in $\boldsymbol{R}^*$ are taken as the mean of the vMF distribution.
2. Input the pair of images into the model to obtain $\hat{\Gamma}_{a,i}(\hat{r}_1,\hat{r}_2,\hat{r}_3)(i=1,2)$,$\hat{\Gamma}_f(\hat{r}_1,\hat{r}_2,\hat{r}_3)$ and $\hat{\boldsymbol{T}}$.
3. According to Eq (9), calculate the mean of $\hat{\Gamma}_{a,i}(\hat{r}_1,\hat{r}_2,\hat{r}_3)(i=1,2)$ and $\hat{\Gamma}_f(\hat{r}_1,\hat{r}_2,\hat{r}_3)$, to obtain $\hat{\boldsymbol{R}}_{a,i}(i=1,2)$ and $\hat{\boldsymbol{R}}_f$.
4. Substitute $\hat{\boldsymbol{R}}_{a,i}(i=1,2)$, $\hat{\boldsymbol{R}}_f$ and $\boldsymbol{R}^*$ into Eq. (13) to obtain $L_{\text{dir}}\left(\hat{\boldsymbol{R}}_{a,i},\boldsymbol{R}^*\right)(i=1,2)$ and $L_{\text{dir}}\left(\hat{\boldsymbol{R}}_f,\boldsymbol{R}^*\right)$.
5. Substitute $\hat{\Gamma}_{a,i}(\hat{r}_1,\hat{r}_2,\hat{r}_3)(i=1,2)$, $\hat{\Gamma}_f(\hat{r}_1,\hat{r}_2,\hat{r}_3)$ and Γ^* into Eq. (14) to obtain $L_{\text{dis}}\left(\hat{\Gamma}_{a,i},\Gamma^*\right)(i=1,2)$ and $L_{\text{dis}}\left(\hat{\Gamma}_f,\Gamma^*\right)$.
6. Obtain $L_{\text{vMF}}\left(\hat{\boldsymbol{R}}_f,\hat{\Gamma}_f\right)$ and $L_{\text{vMF}}\left(\hat{\boldsymbol{R}}_{a,i},\hat{\Gamma}_{a,i}\right)$ as Eq. (15).
7. $L_{\text{att}}=L_{\text{vMF}}\left(\hat{\boldsymbol{R}}_f,\hat{\Gamma}_f\right)+\sum_{i=1}^{2}w_iL_{\text{vMF}}\left(\hat{\boldsymbol{R}}_{a,i},\hat{\Gamma}_{a,i}\right),w_i\in(0,1)$
8. $L_{\text{tran}}=\left(\hat{\boldsymbol{T}}-\boldsymbol{T}^*\right)^2$
9. $L_{\text{total}}=L_{\text{tran}}+L_{\text{att}}$

3. Experiment and analysis

3.1. Dataset

Pairs of UAV and satellite images from 3 different cities across England are created for our dataset. For our standard dataset, the training, validation, and test sets all are images from Oxford and Cambridge. For our generalized dataset, the training and validation sets are from Oxford and Cambridge while the test sets are from London.

We use Google Earth Studio to simulate small UAV images with various poses firstly. For each of the cities, we define several central points around which the flight curves are set, and record the flight video with 30 frames per second. The flight altitude is about 80–120 m. Each UAV image is defined by $\{L,\lambda,h,\theta,\gamma,\psi\}$, where L,λ and h represent the latitude, longitude and altitude respectively, the pitch $\theta\in[-50^\circ,0^\circ]$, the roll $\gamma\in[-10^\circ,10^\circ]$, and the head $\psi\in[-180^\circ,180^\circ]$.

Secondly, we extract a corresponding satellite image of the chosen cities from Google Map. The altitude of the center point of the satellite image is 1000 m, and the longitude and latitude are the same as that of the central points of the UAV flight curves. All the satellite images cover an area of about 2 km × 1 km, with 0° heading, i.e., are aligned with the North direction. All the pairs of images are scaled to a fixed size of 224 × 224 pixels.

3.2. Evaluation metrics

In this paper, the average absolute errors of the translation and Euler angle are reported to evaluate the estimation accuracy of the proposed method. Additionally, the proposed Pd/Re-Net with AGTFM model directly outputs the direction vectors in the rotation matrix, and thus the geodesic error [36] is also reported to evaluate the estimation accuracy of the unit vectors intuitively.

The geodesic error of the rotation matrix between the estimation and the ground truth is:

$$\delta\phi = \arccos\frac{\mathrm{trace}\left(\hat{\boldsymbol{R}}(\boldsymbol{R}^*)^{\mathrm{T}}\right) - 1}{2} \tag{16}$$

3.3. Discussion on the design of proposed Pd/Re-Net

The GoogleNet is initialized with the pre-trained weight on the Cambridge Landmark dataset. [28] Then, the Pd/Re-Net with AGTFM model is trained with loss weights $\alpha = 8 \times 10^7$, $w_1 = w_2 = 0.3$. SGDM optimizer [37] with a learning rate of 0.001 is chosen. The batch size is 20. And the ground truth distributions are generated by the von Mises-Fisher distribution with $\kappa = 10$.

3.3.1. Comparison with state-of-the-art methods

To verify the correctness of our continuity analysis and the advantage of the proposed model, we compare the performance of multiple variants of our model on both the standard dataset and the generalized dataset. For analyzing the effectiveness of our approach to attitude estimation, we focus on the two main questions: "rotation matrix versus other attitude representations" and "probability distribution versus direct regression". Therefore, the regression baselines include networks that directly regress the rotation matrix, equivalent rotation vector, Euler angle, and quaternion, which are represented as Re-Net-R, Re-Net-Rv, Re-Net-E, and Re-Net-Q respectively. All the regression baselines are constructed by changing the parameters of the last fully-connected layer in Rpnet. [38] Probabilistic baselines include models that estimate the probability distribution of equivalent rotation vectors, Euler angles, and quaternions respectively, which are represented as Pd-Net-Rv, Pd-Net-E, and Pd-Net-Q correspondingly. The probability distribution model of the equivalent rotation vector is built on the basis of DirectionNet-R [27] by adding a parallel branch to estimate the two-norm of the equivalent rotation vector; 3D-RCNN [39] estimates Euler angles with a discrete von Mises distribution over the quantized circle; the probability distribution model of the unit quaternions is built at resolution 16^3 over the 4D unit sphere as well. All the above baselines share the same image encoder architecture and are without the proposed attention-guided triple fusion mechanism.

As illustrated in Tables 1–2, our Pd/Re-Net model outperforms all baselines on each metric on both the standard and the generalized dataset. Besides, both show the same following laws: Prediction based on the probability distribution is consistently better for each angle when we compare the Pd/Re-Net model with the Re-Net-R model, which benefits from the accurate estimation of attitude uncertainty. However, the same improvement is not observed apparently when the attitude is represented by the discontinuous equivalent rotation vector or the quaternion. The circular normal distribution does not help at all when represented by the Euler angle, even performed worse. These results confirm that the continuity analysis and the selection of attitude representations are important.

For translation estimation evaluation, we focus on the main question of "probability distribution versus direct regression". DirectionNet-T [27] predicts the translations without scale from the rotation-free image pairs. We use the Direction-Net-T model without de-rotating as the baseline, which added an additional branch layer to estimate the scale and is represented by Re-Net-T. Tables 3–4 report the mean absolute errors of the translation estimates on the standard and generalized datasets, respectively. It is concluded that the probability distribution layer does not play an important role in the improvement of translation estimation accuracy, especially on the generalized dataset. Considering that the probability distribution layer will increase the number of model parameters and computational complexity, the proposed Pd/Re-Net model, which predicts translation by regression, can provide a good trade-off between improving performance and increasing model complexity.

In addition, it can be seen that for the rotation matrix representation with good continuity, the results on the standard dataset are always better than those on the generalized dataset; however, for the other attitude representations, due to the poor continuity, the appearance of intermittent points will affect the fitting ability of the network and some outliers will appear. Therefore, the results on the standard test set are sometimes slightly larger than those on the generalized test set for the other attitude representation methods. For the translation prediction, the accuracy on the standard dataset is always superior to that on the generalized dataset by comparing Tables 3–4. As a result, the suitable input parameters with good continuity can better exploit the strong fitting ability of neural networks, which also proves the significance of the analysis of the various attitude representations.

However, "Satellite-UAV" image pairs, in reality, are often significantly different as discussed in Section 1, posing a huge challenge to the UAV absolute pose estimation based on the pre-loaded satellite image. And the usage of simulated datasets with wide availability is one of the main approaches to address that. Therefore, in the following discussions and analysis, we will mainly focus on the performance of the proposed model on generalized datasets.

Table 1 Comparison of attitude estimation by different methods on standard dataset.

Model name	Estimation parameter	Attitude representation	$\delta\theta(^\circ)$	$\delta\gamma(^\circ)$	$\delta\psi(^\circ)$
Pd/Re-Net	Probability distribution	Rotation matrix	1.34	2.32	5.78
Pd-Net-Rv		Equivalent rotation vector	15.45	14.00	59.82
Pd-Net-E		Euler angles	7.24	145.18	121.11
Pd-Net-Q		Quaternion	15.29	3.75	62.28
Re-Net-R	Regression	Rotation matrix	1.96	2.8	9.44
Re-Net-Rv		Equivalent rotation vector	8.65	5.63	90.79
Re-Net E		Euler angles	5.22	40.64	60.79
Re-Net-Q		Quaternion	2.08	5.94	15.9

Table 2 Comparison of attitude estimation by different methods on generalized dataset.

Model name	Estimation parameter	Attitude representation	$\delta\theta(°)$	$\delta\gamma(°)$	$\delta\psi(°)$
Pd/Re-Net	Probability distribution	Rotation matrix	4.14	3.16	49.95
Pd/Re-Net-Rv		Equivalent rotation vector	8.31	7.06	55.94
Pd/Re-Net-E		Euler angles	12.57	253.81	102.88
Pd/Re-Net-Q		Quaternion	12.48	6.01	90.00
Re-Net-R	Regression	Rotation matrix	7.5	7.35	65.11
Re-Net-Rv		Equivalent rotation vector	18.26	6.44	88.87
Re-Net E		Euler angles	4.54	107.05	86.90
Re-Net-Q		Quaternion	6.77	8.69	88.65

Table 3 Comparison of translation estimation by different methods on standard dataset.

Model	δE(m)	δN(m)	δU(m)
Pd/Re-Net	9.28	10.35	6.41
Re-Net-T	5.46	1.6	9.22

Table 4 Comparison of translation estimation by different methods on generalized dataset.

Model	δE(m)	δN(m)	δU(m)
Pd/Re-Net	87.94	42.31	20.53
Re-Net-T	81.82	47.10	9.88

3.3.2. Robustness analysis

Figs. 5–6 show the performance of the proposed Pd/Re-Net model with different satellite image scales and view angles, compared with the Re-Net-R method. It can be seen that as the satellite image shooting heights increase, the larger the size of the satellite image, the more interference information irrelevant to the UAV image, and the greater the error of attitude estimation using the regression method, while with the increase of the view difference between the "UAV-satellite" image pairs, the shared information becomes less, which also brings trouble to the regression model. However, the proposed Pd/Re-Net method consistently exhibits higher estimation accuracy and stability, which demonstrates that the Pd/Re-Net model quantifies the uncertainty of pose predictions correctly and is able to model ambiguous and challenging cases.

It should be extra explained that the roll axis and yaw axis of the UAV are aligned when the view difference is within the range of 0° to –20°, limited by Google Earth Studio. Hence neither the probability distribution model nor the regression model can achieve highly accurate estimation within this range.

3.3.3. Discussion on the choice of concentration parameter

Fig. 7 compares the effect of different values of concentration parameter κ in the Pd/Re-Net model on attitude estimation accuracy. The highest prediction accuracy is obtained for all the metrics when $\kappa = 10$. As shown in Fig. 3, when $\kappa = 1$, the directional samples are greatly scattered, which is a challenge for the learning ability of the neural network at the resolution of 64^2 set in this paper; when $\kappa = 50$, the samples are concentrated near the mean, making the model susceptible to outliers and not robust to complicated scenarios, thus causing a decrease in accuracy. Therefore, $\kappa = 10$ is reasonable under the design of the proposed Pd/Re-Net model.

3.4. Visualization and ablation study

In Fig. 8, we visualize the feature maps extracted by the proposed attention-guided triple fusion GoogleNet as described in Section 2.2.3. Given a pair of "satellite-UAV" images in Figs. 8(a)–(b), we generate three levels of feature maps from Pd/Re-Net and Pd/Re-Net with AGTFM, respectively in Figs. 8(c)–(h) and Figs. 8(i)–(n). It is illustrated that the feature maps in the first level extract more pixel-level detail information, such as the texture, corner, and other relevant information of the point and its adjacent points, those in the last level extract target-level global semantic information, and

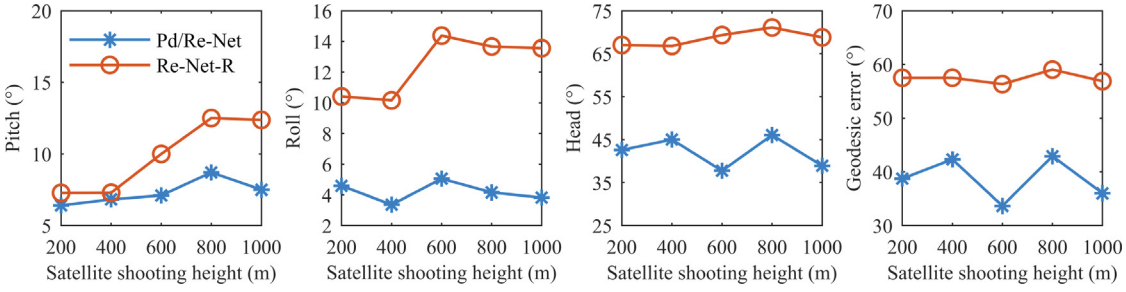


Fig. 5 Estimation error under various satellite image shooting heights.

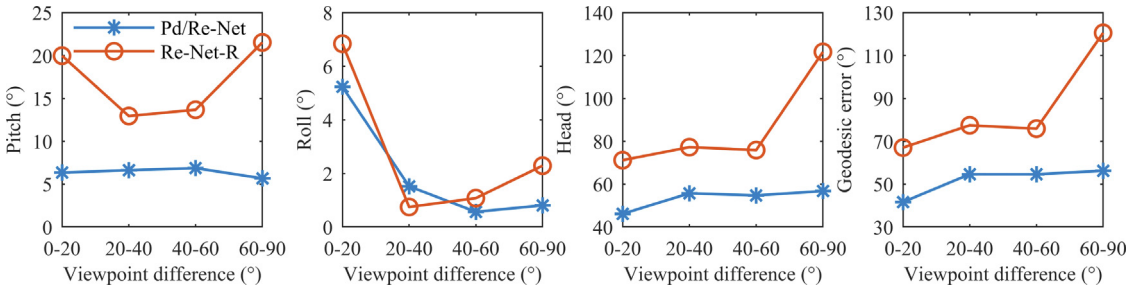


Fig. 6 Estimation error under various viewpoint differences between "satellite-UAV" image pairs.

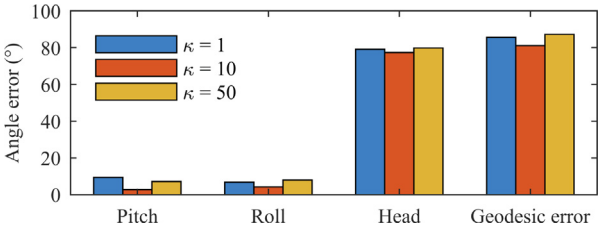


Fig. 7 Attitude estimation error under various concentration parameters.

those in the second level extract both. Additionally, the feature maps extracted by the model with the attention mechanism better align with the semantics of the images and provide more high-level information that highlights informative regions. For example, the big lawns in Fig. 8(b) are taken as visually salience features, which are more significant in location in Fig. 8(l).

Furthermore, the ablation study is conducted to quantitatively evaluate the effectiveness of the proposed attention-guided triple fusion mechanism, as illustrated in Fig. 9, where "a-f" stands for the model with an attention-guided triple fusion module; "f" is the model optimized by the triple fusion module only, where the three image embeddings at different levels extracted from GoogleNet are used to estimate the attitude; "a" is the model optimized by an attention module only, where the image embedding 3 extracted from the GoogleNet improved by Senet is used to estimate the attitude; "/" is the original Pd/Re-Net model without any extra module. We observe that the attention-only mechanism can improve the estimation accuracy to a certain extent by highlighting the key regions. Meanwhile, the fusion-only module fuses the pixel-level details of both "key regions" and "unimportant regions" simultaneously, interfering with the model, resulting in little help on attitude estimation, especially in the roll and yaw. It is concluded that compared with the GoogleNet, our proposed network can not only focus on the key regions but also enlarge local details, amplifying valuable information and resulting in more accurate results.

3.5. Flight experiment

In order to verify the robustness of the proposed Pd/Re-Net with AGTFM model in real scene, in January 2023, we used DJI Matrice 600 Pro with a forward-facing camera, to take UAV images of an industrial area in the suburbs of Beijing. The flight height is about 70–90 m. Three sets of images are captured with camera pitch angles of 0°, –10°, and –50°, examples of which are shown in Figs. 10(a) - (c). The corresponding satellite remote sensing image over the industrial area is acquired from Google Map, covering an area of about 2.5 k

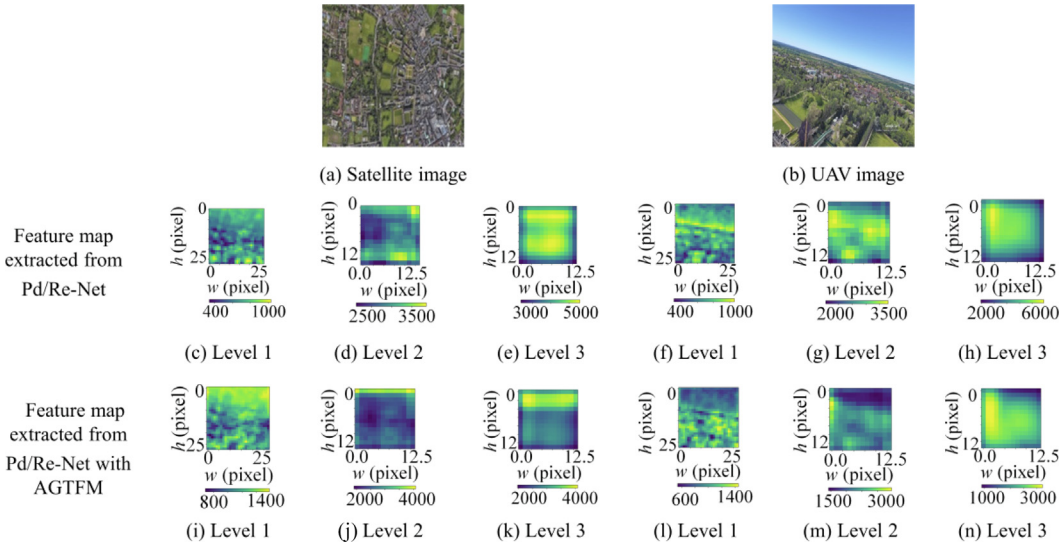


Fig. 8 Visualization of feature maps extracted from Pd/Re-Net and Pd/Re-Net with AGTFM.

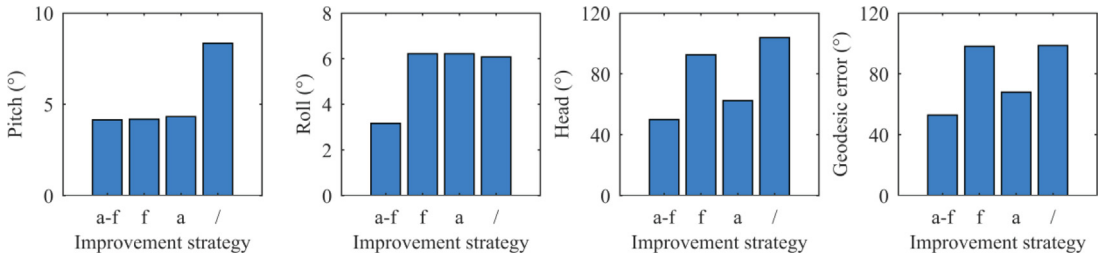
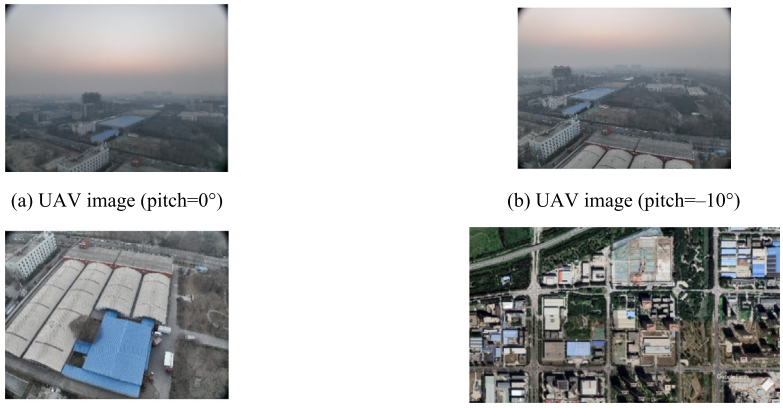


Fig. 9 Ablation study results.



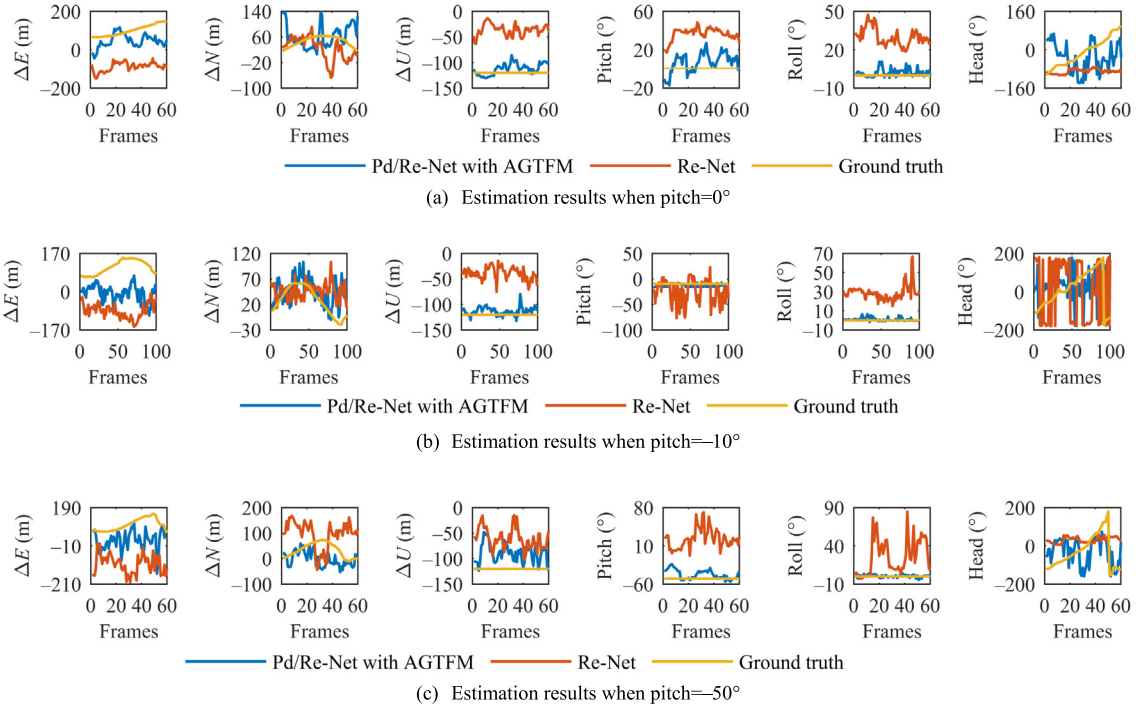
(a) UAV image (pitch=0°)

(b) UAV image (pitch=–10°)

(c) UAV image (pitch=–50°)

(d) Satellite image

Fig. 10 Examples in flight test dataset.



(a) Estimation results when pitch=0°

(b) Estimation results when pitch=–10°

(c) Estimation results when pitch=–50°

Fig. 11 Flight test results estimated by Pd/Re-Net with AGTFM and Re-Net.

Table 5 Comparison of pose estimation errors by different methods on flight test dataset.

Pitch	Model	δE(m)	δN(m)	δU(m)	$\delta\theta$(°)	$\delta\gamma$(°)	$\delta\psi$(°)	$\delta\phi$(°)
0°	Pd/Re-Net with AGTFM	74.90	35.71	10.13	10.26	2.57	87.49	85.65
	Re-Net	187.60	42.38	82.11	33.57	33.25	88.55	124.91
–10°	Pd/Re-Net with AGTFM	111.62	25.29	8.34	4.14	1.28	75.91	74.87
	Re-Net	191.60	29.00	80.26	19.50	28.89	105.47	80.33
–50°	Pd/Re-Net with AGTFM	79.76	40.22	30.42	8.66	1.65	73.51	72.47
	Re-Net	205.48	89.45	59.40	73.15	26.59	91.91	89.92

Table 6 Comparison of computational cost pose estimation by different methods on flight test dataset.

Model	Model size (MB)	MFLOPs
Re-Net	65.60	24.43
Pd/Re-Net with AGTFM	95.60	90.96
Intel Movidius Myriad 2 VPU	4096	100,000

m × 1.2 km, which was updated in April 2021, as shown in Fig. 10(d). It is obvious in Fig. 10 that the satellite image differs considerably from the real-time images in terms of illumination, movement of small objects (e.g., vehicles and trailers), changes in large structures (e.g., building additions and demolitions), trees, etc., due to the weather, different shooting times, and seasons. Consequently, the flight test dataset is more complex and challenging than the simulated one. The estimation results are shown in Fig. 11 and Table 5. It can be seen that the direct regression network has a sharp increase in the pose estimation error in the real flight scenario, while our network shows better accuracy and robustness. In addition, since we did not use any temporal continuous information or inertial information, the results oscillate severely in Fig. 11, which could be improved by building a vision/inertial integrated navigation model.

3.6. Computational cost

In Table 6, we provide a detailed computation comparison between the proposed Pd/Re-Net with the AGTFM method and the traditional method, Re-Net, which is an important algorithmic aspect when deployed on the Vision Processing Units (VPU) of the small UAV. As indicated in Tables 5–6, although the designed Pd/Re-Net with AGTFM model adds an additional 272.2% computational complexity and 45.73% number of parameters, the average improvement in translation estimation accuracy is 244.85% and attitude estimation accuracy is 528.96% during the actual flight, which also exhibits higher robustness facing the challenging scenarios with strong viewing angles and satellite scale differences. We believe that the small additional computational cost incurred by the Pd/Re-Net with the AGTFM model is justified by its contribution to model performance. In order to evaluate the practical application capability of our model, the performance parameters of the VPU used in current small consumer-grade UAVs are also given in Table 6, as exemplified by the Intel Movidius Myriad 2 VPU employed by the DJI Spark mini-UAV. It is concluded that the proposed model is fully deployable in this VPU, even without any lightweight design. As a result, the proposed Pd/Re-Net with the AGTFM model is suitable for a wide range of application platforms and maintains a good balance between performance gains and increased model complexity.

4. Conclusions

In order to solve the difficulty of relative pose estimation between "UAV-satellite" image pairs with large viewpoints and scale differences, and to improve the autonomous navigation capability and survivability of small UAVs in a GNSS-denied environment, this paper proposes a probability distribution/regression integrated deep model with the attention-guided triple fusion mechanism. The following conclusions can be drawn from this study:

(1) In this paper, a deep model combining probability distribution and regression is developed for accurate estimation of the von Mises-Fisher distribution in attitude space and the 3D vectors in translation space. The attitude is represented by a rotation matrix with good continuity, which significantly improves the fitting ability of the neural network compared to other representations, such as equivalent rotation vectors, Euler angles, and quaternions. Then the built model learns the von Mises-Fisher distribution for each direction vector in the rotation matrix and exhibits higher estimation accuracy compared to the direct regression approach. Meanwhile, the translation is estimated by direct regression, which reduces the complexity of the network model as well as maintains a high estimation accuracy of the network.
(2) This paper proposes an attention-guided triple fusion module, which enables the feature extraction module to focus more on the beneficial region to the relative pose estimation and amplifies the pixel details of that region, thus improving the estimation accuracy of the model.
(3) The performance of the proposed method is validated by simulated datasets and flight tests, indicating that our proposed Pd/Re-Net with AGTFM model consistently shows better estimation accuracy, robustness, and generalization ability compared to direct regression networks.

In the future, we consider adding inertial sensors to further improve the estimation accuracy and robustness of the system.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 61973033) and the Chongqing Natural Science Foundation, China (No. cstc2021jcyj-msxmX0737).

References

1. Zhou LY, Zhao XT, Guan X, et al. Robust trajectory planning for UAV communication systems in the presence of jammers. *Chin J Aeronaut* 2022;**35**(10):265–74.
2. Hu ZJ, Gao XG, Wan KF, et al. Imaginary filtered hindsight experience replay for UAV tracking dynamic targets in large-scale unknown environments. *Chin J Aeronaut* 2023;**36**(5):377–91.
3. Tang X, Wang W, He HL, et al. Energy-efficient data collection for UAV-assisted IoT: joint trajectory and resource optimization. *Chin J Aeronaut* 2022;**35**(9):95–105.
4. Atif M, Ahmad R, Ahmad W, et al. UAV-assisted wireless localization for search and rescue. *IEEE Syst J* 2021;**15**(3):3261–72.
5. Zhao XR, Yang RN, Zhang Y, et al. Deep reinforcement learning for intelligent dual-UAV reconnaissance mission planning. *Electronics* 2022;**11**(13):2031.
6. Zhen ZY, Wen LD, Wang BL, et al. Improved contract network protocol algorithm based cooperative target allocation of heterogeneous UAV swarm. *Aerosp Sci Technol* 2021;**119**:107054.
7. Zhao CH, Yang ZY, Cheng XR, et al. SINS/GNSS integrated navigation system based on maximum versoria filter. *Chin J Aeronaut* 2022;**35**(8):168–78.
8. Gao BB, Hu GG, Zhang L, et al. Cubature Kalman filter with closed-loop covariance feedback control for integrated INS/GNSS navigation. *Chin J Aeronaut* 2023;**36**(5):363–76.
9. Fang W, Zheng LY. Rapid and robust initialization for monocular visual inertial navigation within multi-state Kalman filter. *Chin J Aeronaut* 2018;**31**(1):148–60.
10. Goforth H, Lucey S. GPS-denied UAV localization using pre-existing satellite imagery. *2019 international conference on robotics and automation (ICRA)*; 2019 May 20-24; Montreal, Canada. Piscataway: IEEE Press; 2019. p.2974–80.
11. Mantelli M, Pittol D, Neuland R, et al. A novel measurement model based on abBRIEF for global localization of a UAV over satellite images. *Robotics Auton Syst* 2019;**112**:304–19.
12. Bianchi M, Barfoot TD. UAV localization using auto-encoded satellite images. *IEEE Robotics Autom Lett* 2021;**6**(2):1761–8.
13. Wu SB, Du C, Chen H, et al. Coarse-to-fine UAV image geolocalizationusing multi-stage lucas-kanade networks. *2021 2nd information communication technologies conference (ICTC)*; 2021 May 7-9; Nanjing, China. Piscataway: IEEE Press; 2021. p. 220–4.
14. Sui HG, Li JJ, Lei JF, et al. A fast and robust heterologous image matching method for visual GEO-localization of low-altitude UAVs. *Remote Sens* 2022;**14**(22):5879.
15. Kinnari J, Verdoja F, Kyrki V. GNSS-denied geolocalization of UAVs by visual matching of onboard camera images with orthophotos. *2021 20th international conference on advanced robotics (ICAR)*; 2021 Dec 6-10; Ljubljana, Slovenia. Piscataway: IEEE Press; 2022. p. 555–62.
16. Kinnari J, Verdoja F, Kyrki V. Season-invariant GNSS-denied visual localization for UAVs. *IEEE Robotics Autom Lett* 2022;**7**(4):10232–9.
17. Tian XY, Shao J, Ouyang DQ, et al. Uav-satellite view synthesis for cross-view GEO-localization. *IEEE Trans Circuits Syst Video Technol* 2021;**32**(7):4804–15.
18. Shi YJ, Li HD. Beyond cross-view image retrieval: Highly accurate vehicle localization using satellite image. *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*; 2022 Jun 19-24; New Orleans, USA. Piscataway: IEEE Press; 2022. p. 17010–20.
19. Zeng ZL, Wang Z, Yang F, et al. GEO-localization via ground-to-satellite cross-view image retrieval. *IEEE Trans Multimed* 2023;**25**:2176–88.
20. Xu Y, Zhang J, Brownjohn J. An accurate and distraction-free vision-based structural displacement measurement method integrating Siamese network based tracker and correlation-based template matching. *Measurement* 2021;**179**:109506.
21. Shetty A, Gao GX. UAV pose estimation using cross-view geolocalization with satellite imagery. *2019 international conference on robotics and automation (ICRA)*; 2019 May 20-24; Montreal, Canada. Piscataway: IEEE Press; 2019. p. 1827–33.
22. Ding LR, Zhou J, Meng LX, et al. A practical cross-view image matching method between UAV and satellite for UAV-based GEO-localization. *Remote Sens* 2020;**13**(1):47.
23. Chen SX, Wu XY, Mueller MW, et al. Real-time geo-localization using satellite imagery and topography for unmanned aerial vehicles. *2021 IEEE/RSJ international conference on intelligent robots and systems (IROS)*; 2021 Sep 27–Oct 1; Prague, Czech Republic. Piscataway: IEEE Press; 2021. p. 2275–81.
24. Bauza M, Valls E, Lim B, et al. Tactile object pose estimation from the first touch with geometric contact rendering. *Proceedings of the 2020 conference on robot learning*. 2021. p. 1015–29.
25. Rempe D, Birdal T, Hertzmann A, et al. Humor: 3D human motion model for robust pose estimation. *2021 IEEE/CVF international conference on computer vision (ICCV)*; 2021 Oct 11-17. Piscataway: IEEE Press; 2021. p. 11488–99.
26. Liu H, Fang S, Zhang ZL, et al. MFDNet: Collaborative poses perception and matrix fisher distribution for head pose estimation. *IEEE Trans Multimed* 2022;**24**:2449–60.
27. Chen KF, Snavely N, Makadia A. Wide-baseline relative camera pose estimation with directional learning. *2021 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*; 2021 Jun 19-25. Piscataway: IEEE Press; 2021. p. 3258–68.
28. Kendall A, Grimes M, Cipolla R. PoseNet: A convolutional network for real-time 6-DOF camera relocalization. *2015 IEEE international conference on computer vision (ICCV)*; 2015 Dec 11-18; Santiago, Chile. Piscataway: IEEE Press; 2016. p. 2938–46.
29. Khan RU, Zhang XS, Kumar R. Analysis of ResNet and GoogleNet models for malware detection. *J Comput Virol Hacking Tech* 2019;**15**(1):29–37.
30. Banerjee A, Dhillon IS, Ghosh J, et al. Clustering on the unit hypersphere using von Mises-Fisher distributions. *J Mach Learn Res* 2005;**6**(9):1345–82.
31. Hillen T, Painter KJ, Swan AC, et al. Moments of von Mises and Fisher distributions and applications. *Math Biosci Eng* 2017;**14**(3):673–94.
32. Hornik K, Grün B. movMF: An R package for fitting mixtures of von Mises-Fisher distributions. *J Stat Softw* 2014;**58**(10):1–31.
33. González Á. Measurement of areas on a sphere using Fibonacci and latitude–longitude lattices. *Math Geosci* 2010;**42**:49–64.
34. Boukaram WH, Turkiyyah G, Ltaief H, et al. Batched QR and SVD algorithms on GPUs with applications in hierarchical matrix compression. *Parallel Comput* 2018;**74**:19–33.
35. Hu J, Shen L, Sun G. Squeeze-and-excitation networks. *2018 IEEE/CVF conference on computer vision and pattern recognition*; 2018 June 18-22; Salt Lake City, USA. Piscataway: IEEE Press; 2018. p. 7132–41.
36. Yang Z, Litany O, Birdal T, et al. Continuous geodesic convolutions for learning on 3D shapes. *2021 IEEE winter conference on*

applications of computer vision (WACV); 2021 Jan 5-9. Piscataway: IEEE Press; 2021. p. 134–44.

37. Dubey SR, Chakraborty S, Roy SK, et al. diffGrad: An optimization method for convolutional neural networks. *IEEE Trans Neural Netw Learn Syst* 2020;**31**(11):4500–11.
38. En S, Lechervy A, Jurie F. RPNet: An end-to-end network for relative camera pose estimation. *European conference on computer vision*; 2018 Sep 8-14; Munich, Germany. Cham: Springer; 2019. p. 738–45.
39. Kundu A, Li Y, Rehg JM. 3D-RCNN: Instance-level 3D object reconstruction via render-and-compare. *2018 IEEE/CVF conference on computer vision and pattern recognition*; 2018 Jun 18-22; Salt Lake City, USA. Piscataway: IEEE Press; 2018. p. 3559–68.