

Vision-Based UAV Localization System in Denial Environments

Dai Ming¹, Huang Jinglin¹, Zhuang Jiedong¹, Lan Wenbo², Cai Yongheng²
and Zheng Enhui^{1*}

¹*Unmanned System Application Technology Research Institute, China Jiliang University,
Baiyang Street, Hangzhou, 310018, Zhejiang, China.

²R&D Department, Aerospace CH UAV Co., Ltd, Haihao Street, Taizhou, 318014,
Zhejiang, China.

*Corresponding author(s). E-mail(s): ehzheng@cjlu.edu.cn;

Contributing authors: s20010802003@cjlu.edu.cn; 1900111330@cjlu.edu.cn;
p1901085206@cjlu.edu.cn; lanwb@caaa.casc; caiyh@caaa.casc;

Abstract

Unmanned Aerial Vehicle (UAV) localization capability is critical in a Global Navigation Satellite System (GNSS) denial environment. The aim of this paper is to investigate the problem of locating the UAV itself through a purely visual approach. This task mainly refers to: matching the corresponding geo-tagged satellite images through the images acquired by the camera when the UAV does not acquire GNSS signals, where the satellite images are the bridge between the UAV images and the location information. However, the sampling points of previous cross-view datasets based on UAVs are discrete in spatial distribution and the inter-class relationships are not established. In the actual process of UAV-localization, the inter-class feature similarity of the proximity position distribution should be small due to the continuity of UAV movement in space. In view of this, this paper have reformulated an intensive dataset for UAV positioning tasks, which is named DenseUAV, aiming to solve the problems caused by spatial distance and scale transformation in practical application scenarios, so as to achieve high-precision UAV-localization in GNSS denial environment. In addition, a new continuum-type evaluation metric named SDM is designed to evaluate the accuracy of model matching by exploiting the continuum of UAVs in space. Specifically, with the ideas of siamese networks and metric learning, a transformer-based baseline was constructed to enhance the capture of spatially subtle features. Ultimately, a neighbor-search post-processing strategy was proposed to solve the problem of large distance localisation bias.

Keywords: Geo-Localization, Transformer, UAV, Deep Learning

1 Introduction

In the usual scenario, UAV localization relied on satellite systems to provide accurate latitude and longitude position information. However, this

approach need to presuppose that satellite systems operate properly and that GNSS signals are not interfered with or blocked. Once the UAV does not acquire GPS signals, the UAV will not

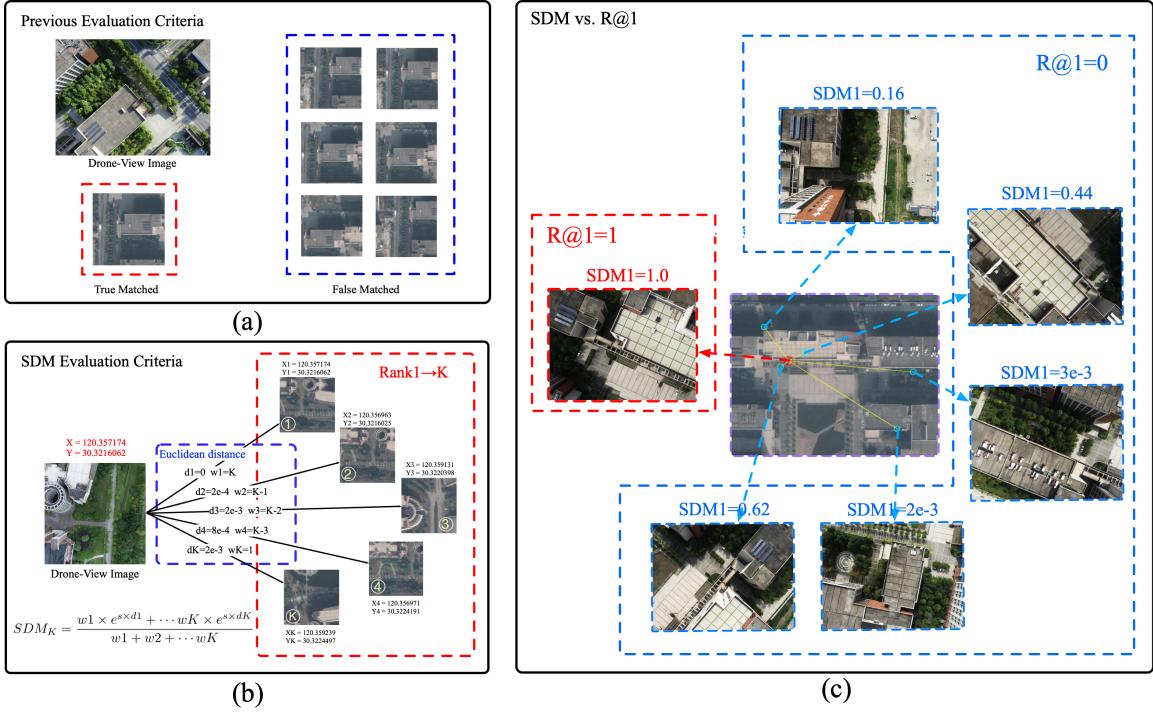


Fig. 1 (a) depicts the criteria for judging the previous evaluation metrics. (b) describes the evaluation criteria of our proposed SDM evaluation metric where d_{1-K} represents the spatial Euclidean distance between the Drone-View Image and the most similar $1-K$ satellite images, and the calculation process is shown in Equation 4, where X represents the longitude and Y represents the latitude. w_{1-K} are weighting factors and s is an amplification factor, set to 5×10^3 in this paper. (c) Compared the evaluation process of R@1 and SDM, for a single sample, the evaluation criterion of R@1 is discrete, i.e. 1 vs. 0. However, for the SDM evaluation metric, the score varies with the spatial distance from ground-truth, with the closer it is the higher the score.

be able to operate properly. Therefore, the ability to locate in denial environment is particularly important. In today's world, UAV has played a more and more pivotal role, and have been used in applications such as military combat, agricultural operations, ground reconnaissance and even civilian aerial photography [Zheng et al, 2020a; Hu and Lee, 2020; Middelberg et al, 2014; McManus et al, 2014]. With the improvement of on-board device performance and the lightweight of hardware, many algorithms such as object tracking, object detection, SLAM and other algorithms [Yu et al, 2020; Ma et al, 2021] have been applied to UAV and generated great social benefits.

The University-1652 [Zheng et al, 2020a] dataset firstly introduces the drone-view to cross-view geo-localization, which regarded drone-view as an intermediate transition view to alleviates the difficulty between matching through ground-view images and satellite-view images. Meanwhile, they have also proposed *drone-view target localization*

and *drone navigation* tasks to extend the application of drone in geo-localization. As a successor, we still explore the subject of drones. The difference, however, is that this paper aims at solving the problem of locating the UAV itself and the implementation is broadly consistent with the geo-localization task, in which the drone-view image is used to retrieve the most similar geo-tagged satellite image in the satellite database to obtain the current location information of the UAV. In a broad sense, The University-1652 locates the building in view, while we locate the device that captures the object such as a camera, whose positioning goal is in the same vein as CVUSA [Zhai et al, 2017a] and CVACT [Liu and Li, 2019a]. The difference is that our approach does not use panoramic images as the images of the drone-view, otherwise we use a simpler solution, i.e., the UAV camera is vertical ground down to solve the view deviation problem, which is an advantage that the ground camera does not have. Another notable

feature is that our dataset is densely collected, which means that there is an overlap between adjacent images, which was designed to increase the influence of spatial position deviation and thus achieve high accuracy positioning.

CMC [Zhao et al, 2013; Zheng et al, 2015] and mAP are the mainstream evaluation metrics for image retrieval tasks. They are based on classification to judge accuracy and are very effective when no spatial relationships are introduced, however, when assessing spatial positioning accuracy, we found some significant drawbacks in these evaluation metrics. In the process of UAV positioning, we expect the matched satellite map to be as close to the actual location as possible. First of all, it should be noted that in the inference stage, the satellite gallery database must be dense, which means that the same prominent landmarks (such as buildings, lakes, trees, etc.) will appear in multiple satellite images. We have made an extreme example in Figure 1(a), where the red box is a satellite image that corresponds exactly to the geographic location of the drone-view image, and others in blue box are satellite-view images with a only small position shift. If the CMC and mAP evaluation metrics are used to judge, only the image in the red box is the True-Matched sample and the images in blue box are all False-Matched samples, which is obviously not in line with the actual application of UAV positioning. Based on this, we propose the evaluation metric SDM to change the previous pattern of 0 or 1 by assigning the corresponding score according to the matching distance, the closer the matching position is to the target, the higher the score should be. This design approach is mainly based on our actual application scenarios, where UAVs inevitably have matching errors during the matching process, and a slight offset is an acceptable phenomenon, while an excessive offset will largely undermine the reliability of subsequent positioning.

The development of transformer in computer vision is in full swing. Compared to the pure CNN approach, transformer shows significant advantages with its self-attention mechanism. Taking a closer look at the CNN-based approach, we mainly found two potential problems. (I) Vision-based geo-localization needs to dig out the relevant information between contexts. Images from different domains have positional transformations such as

rotation, scale and offset. Therefore, fully understanding the semantic information of the global context is necessary. However, CNN-based methods mainly focus on small discriminative regions due to Gaussian distribution of effective receptive fields [Vaswani et al, 2017]. Given the limitations of the pure CNN-based methods, the attention modules have been introduced to explore long-range relationships [Wang et al, 2018]. However, most of the methods embed the attention mechanism into the deep convolutional network, which enhances contextual connections to a certain extent. (II) Fine-grained information is very important for the task of retrieval. The down-sampling operations i.e., pooling and stride convolution of the CNN-based method can reduce the resolution of the image, while inevitably destroying the recognizable fine-grained information. In this paper, we propose a Transformer baseline for the task of UAV vision localization. Similar to the cross-view ground-aerial matching task, we use siamese network architecture with shared weights, a fusion of metric learning and representation learning to learn intra-class similarity and inter-class differences. Conventionally, we use feature vectors to characterize the images of drone-view and satellite-view, and euclidean distance to measure the gap between features representation.

Our benchmark provides an end-to-end UAV-localization solution. The complete process is shown in Figure 2. (1) Acquire satellite images from Google Earth with 20-level and perform isometric multi-scale cutting of the satellite maps to build satellite database. (2) The images in satellite database are then passed through our geo-localization network to obtain the corresponding feature representations, and the UAV images undergo the same operation mentioned above. (3) Compute the feature distance between the drone-view image and every image in satellite databases. In order to visualize the process of localization, we use the GPS information that comes from the UAV to do the verification (in the real scenario the UAV does not have GPS, so this is just for visualization). (4) The localization strategies part is used to control the scope of the search space (global search or neighbor-search). (5) Then we can retrieve the satellite images with the closest distance to the drone-view image from the satellite database, and the top1-10 images are chosen to show in this paper. Finally, we visualize some

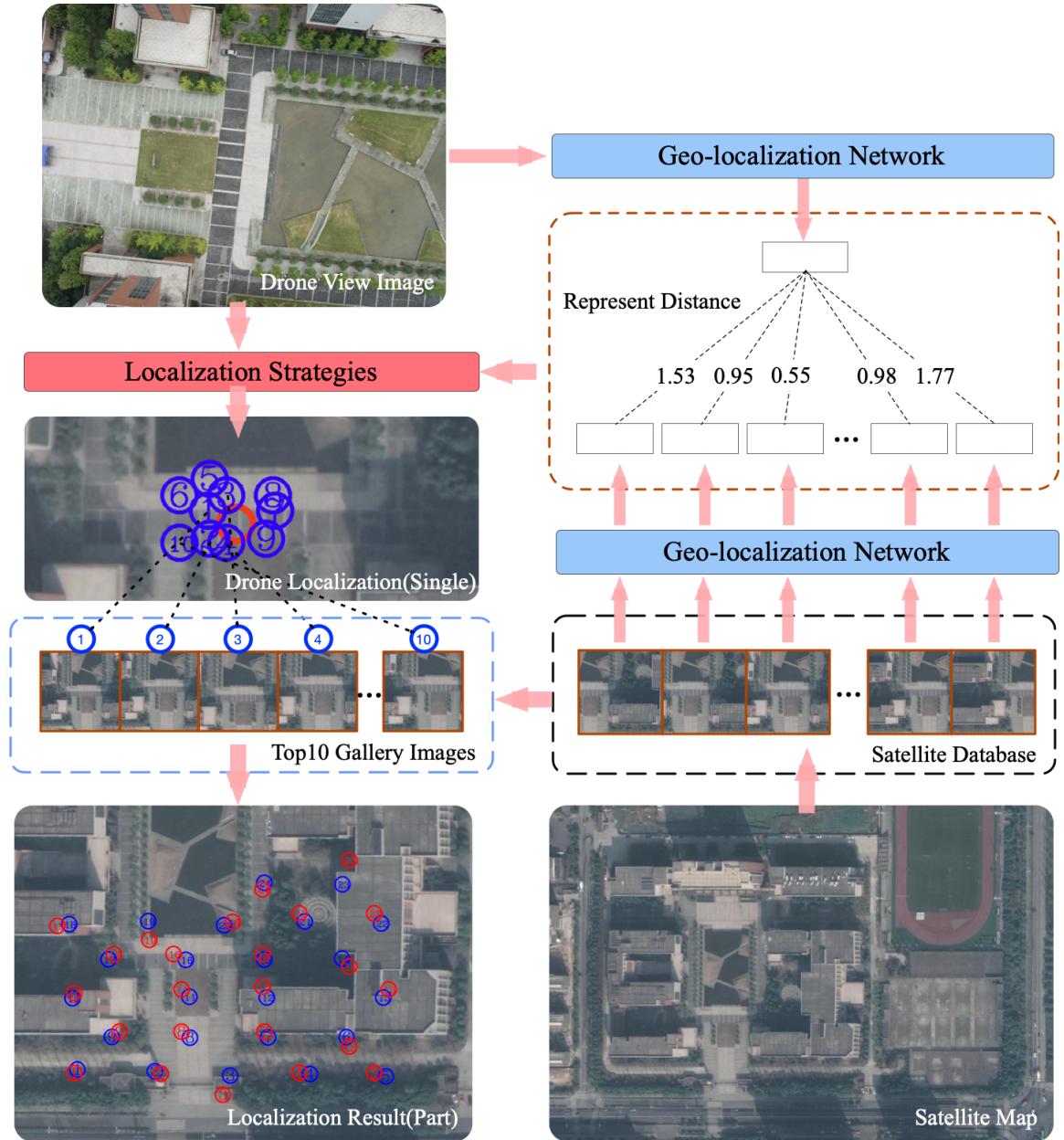


Fig. 2 Diagram of the flow structure of a UAV positioning mission during the execution phase. localization Strategies refer to post-processing operations after model inference, such as Global-Search and Neighbor-Search. blue circles in localization Result (Part) are real UAV locations, red are matched locations, and numbers indicate correspondence.

images taken by UAV in the local map (using top1 satellite image as the matching term).

In the process of actual application, we found that the UAV positioning system based on global searching still has many cases of incorrect matching. One of the most important reasons is that the satellite map is so large that the images cut out from the map are somewhat very similar.

Coupled with the fact that the satellite images themselves are of low clarity and there will also be areas covered by shadows. Thus, the probability of incorrect matching will be greatly increased. An efficient solution to alleviate the above problems is to reduce the search range. Relying on the continuity of object motion in space, we adopt the neighbor-serach strategy to avoid some images

with high similarity but far away from each other. In the implementation, we construct a circle with a constant distance based on the GPS information obtained from the previous match, the radius of the circle is also determined by the flight speed and positioning sampling frequency of the UAV. Using this circle as the search domain, the satellite images within the circle are filtered to construct a new satellite database as the candidate gallery library. This approach not only improves the accuracy but also the speed of retrieval. However, a hyper-parameter M will be introduced and this hyper-parameter needs to be adjusted in combination with the UAV related parameters during the flight of the UAV to achieve better performance.

The contributions in this paper are as follows.

- we propose a dense UAV-based geo-localization dataset, named DenseUAV, which aims to achieve high-precision positioning of drones.
- We propose a new evaluation criterion, SDM, to assess the accuracy of UAV positioning based on the deviation of spatial position.
- Based on the proposed DenseUAV dataset, we constructed a transformer-based baseline to accomplish the UAV localization task. Meanwhile, an improved neighbor-search scheme is proposed to further enhance the reliability of UAV visual localization in the denial environment.

2 Related Work

2.1 Geo-Localization Dataset

Initially, the geo-localization task was proposed to solve the ground-to-aerial matching problem, where ground refers to images captured by users using mobile cameras and aerial refers to images captured by satellites. One of the earliest works [Lin et al, 2015] proposed to leverage the public sources to build image pairs for the ground-view and aerial-view images, which consists of 78k image pairs from two views, i.e., 45° bird view and ground view. Subsequently, [Tian et al, 2017] collected image pairs from the town view for the task of geo-localization, which argued that architecture can provide more notable features for geo-localization. [Vo and Hays, 2016] aimed to determine the location and orientation of a ground-level query image by matching a reference

database of satellite images, and they collected a dataset with one million pairs of street view and overhead images sampled from eleven U.S. cities. Later, two well-known large-scale datasets i.e., CVUSA [Zhai et al, 2017a] and CVACT [Liu and Li, 2019a] made grand entrance. [Zhai et al, 2017a] constructed image pairs from ground-based panoramic images and satellite images, while [Liu and Li, 2019a] added spatial factors to CVUSA, i.e., orientation maps. The latter three mentioned datasets targeted users with photographic devices, while the first two mentioned datasets targeted the buildings in images. Recently, University-1652 [Zheng et al, 2020a] introduced the drone-view into cross-view geo-localization, using the drone-view as a transition view to reduce the difficulty of the matching between ground-view and satellite-view, and regarded the cross-view geo-localization as a retrieval task. In addition, University-1652 proposed two UAV-based subtasks respectively called drone-view target localization and drone navigation, which drove the application of UAVs for cross-view geolocation. The emergence of DenseUAV was inspired by University-1652, the different is that University-1652 focuses on the recognition of buildings and selects buildings with certain features and styles, and the space distribution of these buildings is discrete. And it aims to solves the problem of localization of buildings in images. In contrast, DenseUAV is designed to solve the problem of the localization of UAV itself under the condition of GPS-free.

2.2 Deeply-Learned Feature For geo-localization

Most prior work has viewed geo-localization as an image retrieval task. The most significant characteristics are the bias in perspective between the matched images and the different sources of the images. Thus a critical factor for cross-view geo-localization to determine retrieval accuracy is the ability to overcome the viewpoint-invariant problem between different domains [Long et al, 2017], such as the existence of a 90-degree span between ground and satellite views, where the ground view can result in many objects visible in the satellite view not appearing in the ground view due to factors such as object occlusion.

One line focus on metric learning and improves the ability of the model to discriminate differentiation by optimizing the loss function. Triplet loss [Liu et al, 2017] was proposed to improve the reliability of matching by closing the distance between features of the positive pairs and distancing the distance between features of negative pairs. The contrastive loss, pulling the distance between positive pairs, could further improve the geo-localization results [Hu and Lee, 2020; Workman et al, 2015]. [Zheng et al, 2020b] applied instance loss and verification loss together to optimize the network, and achieve competitive results. [Hu et al, 2018] proposed a weighted soft margin ranking loss, which not only speeds up the training convergence but also improves the retrieval accuracy. [Liu and Li, 2019b] proposed a BNNeck to improve the coordination of ID loss and triplet loss. [Sun et al, 2020] proposed a unified framework to optimize classification learning and pairwise learning.

Another line focus on the spatial information asymmetry between ground-to-aerial matching. [Zhai et al, 2017b] proposed an end-to-end learning approach to minimize the difference between the semantic segmentation extracted directly from the ground image and the semantic segmentation predicted solely based on the aerial image. [Arandjelovic et al, 2016] inserted the NetVLAD layer to express the relationship between features and each cluster in terms of probability values and extract discriminative features. Further, [Liu and Li, 2019b] designed a Siamese network which explicitly encodes the orientation (i.e., spherical directions) of each pixel of the images. Similarly, [Shi et al, 2020b] firstly applied a regular polar transform to warp an aerial image to bring its domain closer to the ground-view panoramic and then add a subsequent spatial-attention mechanism to improve the robustness of feature representation. [Shi et al, 2020a] addressed the cross-view domain gap by applying a polar transform to the aerial images to approximately align the images up to an unknown azimuth angle. [Toker et al, 2021] proposed a GAN structure to create realistic street views from satellite images and localizes the corresponding query street-view simultaneously in an end-to-end manner. [Wang et al, 2021] proposed a square chunking strategy to fully extract useful information from the edges,

and the chunking scheme results in a significant improvement in localization performance.

2.3 Transformer In Vision

The attention mechanism [Vaswani et al, 2017] of the transformer model was first proposed to solve problems in the field of Natural Language Processing. recently, [Dosovitskiy et al, 2020] firstly applied the transformer model to the task of image classification. Subsequently, Transformer has achieved state-of-the-art performance in various domains of vision. DETR [Carion et al, 2020] was the first object detection framework that successfully integrates the transformer as the central building block of the detection pipeline. SETR [Zheng et al, 2021] treated semantic segmentation as a sequence-to-sequence prediction task through a pure transformer. TransGAN [Jiang et al, 2021] built a generator and a discriminator based on two transformer structures. TTSR [Yang et al, 2020] restored the texture information of the image super-resolution result based on the transformer. TransReID [He et al, 2021] applied the transformer to the field of retrieval for the first time and achieved similar results with the CNN-based method.

Also there is a lot of work using Transformer as the backbone network to extract features, DeiT [Touvron et al, 2021] introduces a teacher-student strategy specific for transformers to speed up Vit training without the requirement of large-scale pretraining data. Swin Transformer [Liu et al, 2021] proposed a hierarchical Transformer whose representation is computed with Shifted windows. MobileVit [Mehta and Rastegari, 2021] combine the strengths of CNN and Vit to build a light-weight and low latency network for mobile vision tasks.

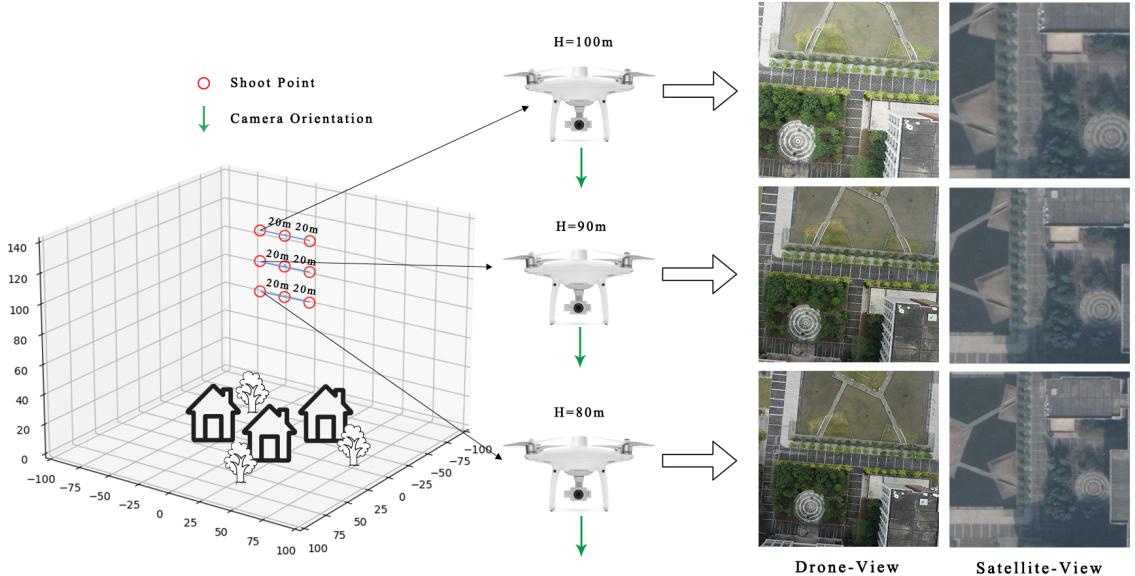
3 Methodology

3.1 Dataset Description

In this paper, we propose a dense UAV-based geo-localization dataset, which differs from previous cross-view geo-localization datasets mainly in the term “dense”. “dense” refers to the fact that the drone-view images in the training phase are dense in spatial distribution, in addition to the fact that the database of satellite to be matched

Table 1 geo-localization dataset information summary table

Datasets	DenseUAV	University-1652	CVUSA	CVACT	Lin et al.	Tian et al.	Vo et al.
#training	2256×6	701×71.64	$35.5k \times 2$	$35.5k \times 2$	$37.5k \times 2$	$15.7k \times 2$	$900k \times 2$
Platform	Drone,Satellite	Drone,Ground,Satellite	Ground,Satellite	Ground,Satellite	Ground,Aerial	Ground,Aerial	Ground,Satellite
#imgs./location	$3 + 3$	$54 + 16.64 + 1$	$1 + 1$	$1 + 1$	$1 + 1$	$1 + 1$	$g 1 + 1$
Target	UAV	Building	User	User	Building	Building	User
Evaluation	SDM	Recall@K & AP	Recall@K	Recall@K	PR curves & AP	PR curves & AP	Recall@K

**Fig. 3** A diagram describing the process of creating the dataset. The red circles represent the sampling locations, with the sampling points spaced 20 metres apart and sampled at 80, 90 and 100 metres respectively. The green arrows represent the orientation of the camera.**Table 2** DenseUAV data volume composition table. Where #classes refers to the number of categories and #universities refers to the number of Universities we sampled.

split	#imgs		#classes	#universities
	drone	satellite		
Training	6768	6768	2256	10
Query	2331	2331	777	4
Gallery	9099	9099	3033	14

in the testing phase is also highly dense in spatial distribution. The ultimate goal of the proposed DenseUAV is to achieve high accuracy positioning

of UAV. To our best knowledge, Another difference is that the DensUAV regards the UAV as the localization target, which has not been proposed and addressed by researchers in the past. As shown in Table 1, we enumerate many elements of the previously proposed dataset for further comparison.

The following three main factors were taken into account in the way of dataset acquisition. First, considering the influence of the height of the UAV flight on the image scale. The UAV images we collected consisted of 3 different heights, 80, 90, and 100 meters (relative to the ground), as shown in Figure 3. The sampling process needs to ensure

that the latitude and longitude of the images sampled at 3 different heights are almost identical. We applied the waypoint action of DJI Plot to achieve this and control the error within 1m. Second, considering the influence of weather and light level on the image quality, we adopt the random weather and random time period sampling method. The random weather mainly includes sunny and cloudy days, and the random time period mainly refers to the random sampling between 6:00 am and 6:00 pm. Third, taking into account the perspective bias caused by the change of UAV camera orientation, we unified the UAV camera with the vertical ground facing downward. In addition, considering the problem of UAV flight orientation, this problem is solved by a kind of rotational data enhancement, which is introduced in Section 4.2. In addition, we set the distance between each sampling point to 20m. There are two purposes for this. On the one hand, we standardize our dataset by this equally spaced sampling, on the other hand, the dataset constructed by dense sampling can enhanced ability of model to learn subtle features.

For the composition of the dataset, we selected 14 universities in Hangzhou, Zhejiang, China as our sampling area due to the large labor and hardware costs for producing the dataset. As shown in Table 2, among them, 6768 drone-view images from 2256 sampling points of 10 universities and an equal amount of satellite images are used for training. In the test set, the query set contains 2331 drone-view images from 777 sampling points of 4 universities and an equal number of satellite images, and the gallery set contains a total of 18198 images from 3033 sampling points of all 14 universities.

3.2 Evaluation Indicators

With recurring content between images in dense datasets and a dense distribution of categories, commonly used evaluation metrics such as Cumulated Matching Characteristics (CMC) curve [Zhao et al, 2013; Zheng et al, 2015] and mean Average Precision (mAP) will no longer be suitable for assessing the accuracy of drone positioning. In the following, we do some specific analysis to explain the reasons for this.

mAP and CMC mainly consider whether the classification is accurate or not. Specifically, these

evaluation criteria are both judging whether the selected image in the gallery belongs to the same category as the image of query. In terms of a single sample, these evaluation metrics simply care whether it is 0 or 1, which we consider as a discrete evaluation indicator. Take rank1 in CMC evaluation index as an example, its expression is shown in Equation 1 and Equation 2.

$$I(l_q, l_i) = \begin{cases} 1 & \text{if } l_q = l_i \\ 0 & \text{if } l_q \neq l_i \end{cases} \quad (1)$$

Where l_q corresponds to the category of query, and l_i corresponds to the category corresponding to the i th image sorted from smallest to largest after calculating the Euclidean distance and ascending the order, which is 1 if it belongs to the same category and 0 if it is not. And the rank1 expression is as follows.

$$\text{rank1} = \frac{1}{\|Q\|} \sum_{q \in Q} I(l_q, l_1) \quad (2)$$

Where Q is the set of all query images and $\|Q\|$ denotes the number of images in the set Q . It can be seen that the rank1 increases when and only when the category of query and the category of the closest image in the gallery are the same, otherwise they are all regarded as False-Matched samples.

However, there is a characteristic in the actual operation of the UAV that the satellite images are spatially continuous and dense, and the gap of feature information contained between adjacent images is small. As shown in Figure 1 (a), if the previous evaluation criteria of CMC and mAP is used, only the satellite images in the red dashed part are correct, $I(l_q, l_1) = 1$, and the satellite images in the blue dashed part are wrong, $I(l_q, l_1) = 0$, although the deviation is very small. In the actual UAV localization task, the UAV is expected to position itself as closer to the real location as possible. Minor positioning deviations is inevitable, however the presence of large deviations is not expected, and the CMC and mAP metrics do not express this difference in spatial distance. In order to increase the judgment of the spatial distance of the UAV, the true GPS coordinate positions of all satellite images and UAV images were recorded. Based on this, an evaluation metric for measuring spatial distance (SDM)

is proposed as an evaluation criterion for UAV localization task. Its expression for a single query sample is as follows.

$$SDM_K = \sum_{i=1}^K \frac{(K-i+1)}{e^{s \times d_i}} \quad (3)$$

$$d_i = \sqrt{(x_q - x_i)^2 + (y_q - y_i)^2} \quad (4)$$

where the numerator $(K-i+1)$ is the weight w , which is presented in Figure 1(b). The weight w is assigned based on the feature distance, and a greater weight is set for the gallery image that is closer to the query feature. K denotes the top K samples in the gallery that are closest to the distance of the feature vector calculated by query. x_q and x_i denote the longitude corresponding to the query and gallery images, respectively, while y_q and y_i denote the latitude. In short, d_i represents the spatial Euclidean distance between two images. S is a scaling factor, and in this paper, S is set to 5×10^3 .

The weights of the SDM are then normalised to give the full formula for the SDM as follows.

$$SDM_K = \frac{\sum_{i=1}^K \frac{(K-i+1)}{e^{s \times d_i}}}{\sum_{i=1}^K (K-i+1)} \quad (5)$$

The interval of SDM metric is distributed between 0 and 1, and the larger SDM value indicates the better localization performance of the model. For a clearer representation, the visualization process of the calculation is presented in Figure 2(b). d_{1-K} represents the spatial distance between query and $rank_{1-K}$ images of the candidate frame.

To illustrate this difference more simply, we compare the SDM1 and R@1 methods in Figure 1(c), where our method is like a continuous evaluation criterion, while R@1 just simply chooses between 0 and 1.

3.3 Benchmark Introduction

Based on the proposed baseline of University-1652, we use small size vision transformer (ViT-S) as the backbone of the network to help the network fully understand the feature relationships between different perspectives. As shown in Table 3. the

performance of the ViT-S-based baseline is much better than that of the Resnet50-based baseline. this is due to the fact that in the task of intensive UAV geo-localization, there are viewpoint biases between images from different sources, making contextual correlation information crucial, and fully understanding the long-range global feature information of images has a significant effect, the shortcoming is that the ViT-S-based model is slightly slower than the resnet50-based model in the inference process, and the inference time is roughly 1.27 times that of the Resnet50.

The overall structure of our baseline is shown in Figure 4, where the input image is cut into 16×16 patches equally, and the information of each patch is compressed by a Linear Projection. Which is composed of a convolution with a kernel of 16×16 and a stride of 16, and expand the number of channels to 768. After that, the dimension is flattened to generate patches, and Each patch plus a Position Embedding. In addition, to express the global features we embed an additional embedding patch for the global feature representation. After passing through the Transformer Layer (The ViT-S used in this paper consists of a total of 8 self-attentive blocks, with the same structure as TransReid [He et al, 2021]), we get the features with the same dimension as the input, and the *cls_token* is regarded as the final feature representation. The output is divided into two parts, one of which converts the 768-dimensional ViT-S output into a 512-dimensional feature vector via a fully-connected layer to calculate Triplet Loss, and the other part goes through a fully-connected layer on this 512-dimensional feature vector for classification.

3.4 UAV Positioning Applications

Another contribution of this paper is that we have made a practical application based on our benchmark. The overall process of UAV localization is actually very simple, and the basic way is to retrieve the closest geo-tagged satellite map from the satellite database through the UAV image, and then obtain the GPS information of the location of the UAV at this time. However, there are some post-processing methods that can simultaneously improve the retrieval speed and accuracy, such as narrowing the retrieval range.

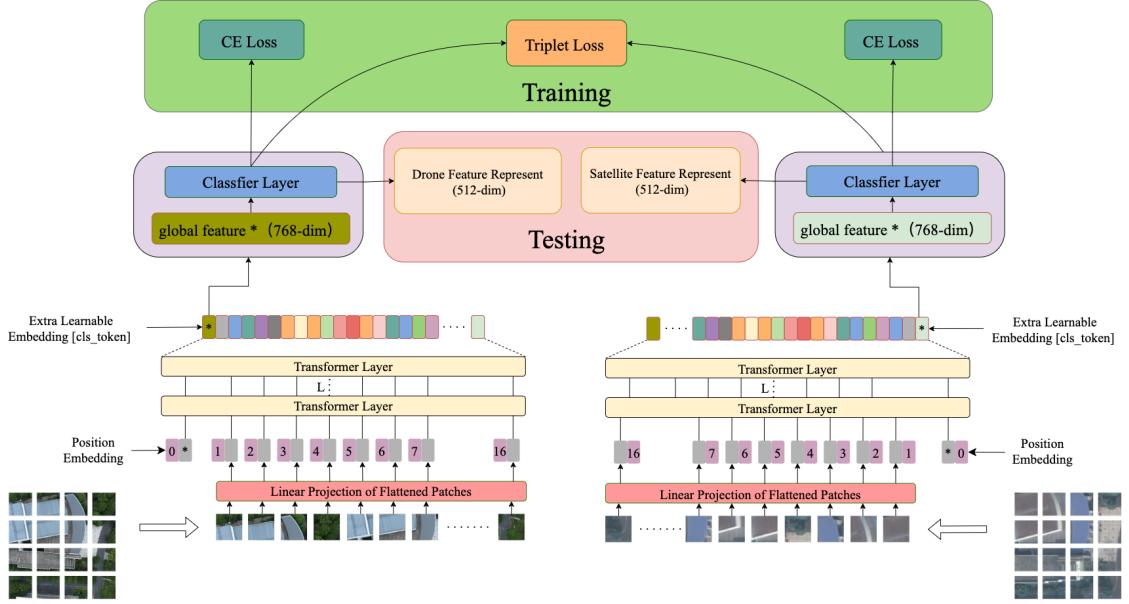


Fig. 4 A diagram of the overall structure of our Baseline which contains both training and testing phases. Classifier Layer contains linear layer, relu, batchnorm1d and dropout. CE Loss represents cross-entropy loss without label-smooth.

Vision-based UAV localization is a fresh approach that can be optimized in terms of localization accuracy and time consumption through many post-processing methods (Corresponds to the positioning strategy in Figure 2), and in this paper we propose the global search method as well as its improved version of the neighbor-search method. The global search method retrieves the whole satellite database, which is downloaded in advance and cut with a sliding window. In order to improve the accuracy of retrieval, this paper adopts the cutting method of multi-scale and repeated area (multi-scale refers to the different sizes of the satellite images, and repeated area refers to the sliding distance of the sliding window is smaller than the window size, and the sliding distance used in the testing process of this paper is 1/4 of the sliding window size), and we have made a visual demonstration in Section 5.1. Although the baseline-based positioning accuracy has achieved good performance, with the increase of the satellite database, the positioning accuracy will be greatly reduced, and there will be a situation of matching to a spatially distant location, producing a large positioning deviation, which is very unfavorable for UAV positioning.

In order to improve the above drawbacks in the global search, we propose a neighbor-search

method to make up for the defects of the global search, which uses the spatial motion continuity of the UAV, takes the previous matching position of the UAV as the center of the circle and takes a suitable radius M to form a circle as the search domain, and the search domain X can be expressed as the following formula.

$$X = \bigcup_{\theta \in H} (d_\theta < M) \quad (6)$$

where H denotes the satellite database, d_θ denotes the distance between the θ th satellite image and the current center position, and M is the search radius. The neighbor-search method eliminates impractical localisation results by reducing the area to be retrieved. However, it is worth noting that the parameter M is required to be adjusted according to the flight speed and sampling retrieval frequency of the UAV. In general, compared to the global search method, the neighbor-search method reduce the scope of retrieval, which can not only reduce the amount of computation but also improve the matching accuracy. However, due to the time difference between the real acquired satellite images, which are historical images, and the UAV images, and the poor clarity of satellite image, the task is still very challenging.

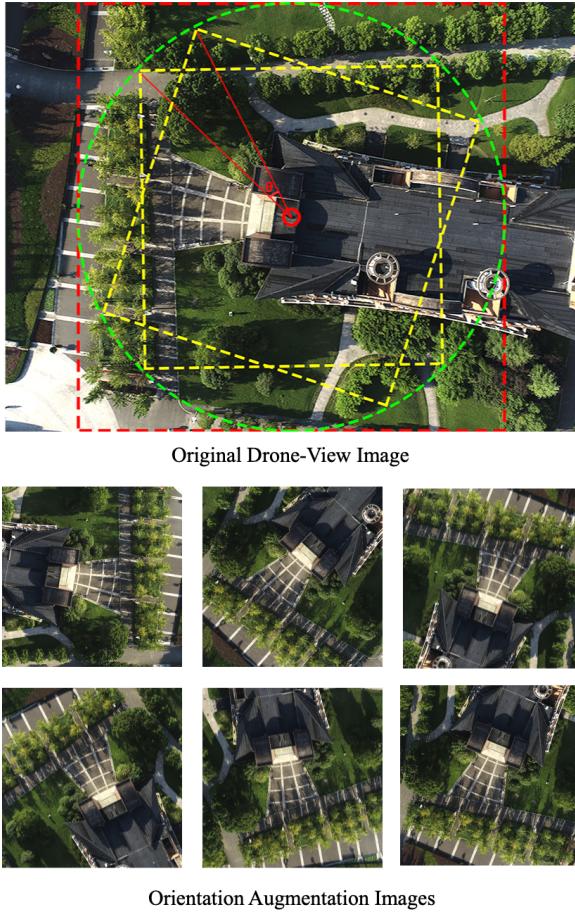


Fig. 5 Schematic diagram of a data enhancement method for enriching the flight orientation information of a UAV. The red dashed line is the largest interior square centred on the centre of the image, the green dashed line is the largest interior circle of that square, and the yellow dashed line represents a number of the largest interior squares of the circle. θ indicates the rotation angle.

4 Ablation Experiments

4.1 Implement Details

We used VitS pre-trained on the ImageNet dataset as the backbone, and added a 512-dimensional fully connected layer and a classification layer after pooling, while removing the original classification branch. The model is trained by stochastic gradient descent with the learning rate is 0.01 for the new-added layers and 0.003 for the rest layers. dropout rate is 0.5. In the training process, the input image size is 256×256 , and some data augmentation methods are applied to expand the data, including color enhancement, horizontal flipping and our proposed rotational method in 4.2.

In the testing phase we use the trained model to extract the features of the images and calculate the Euclidean distance between the features to characterize the similarity between the query and the candidate images in the gallery. The final retrieval results are derived based on ranking of feature distance. It is worth mentioning that for different views, we use only one model for training, which means that all weights are shared.

Table 3 Table of experimental results for UAV flight direction data augmentation. o denotes without our proposed data augmentation, w denotes with.

Backbone	w/o rotate	SDM1	SDM3	SDM5	SDM10
Resnet50	o	0.143	0.125	0.115	0.100
Resnet50	w	0.155	0.136	0.128	0.114
VitS	o	0.454	0.436	0.390	0.305
VitS	w	0.525	0.504	0.449	0.345

4.2 Flight direction data expansion

In the Vision-based UAV localization task, the UAV and satellite images are supposed to correspond one-to-one, but such one-to-one correspondence is not very helpful for model learning, and we expect to create more situations that match real scenarios. Both the flight direction and the flight altitude of the UAV are uncertainties in real scenarios. Intuitively, these two factors affect the rotation and scale of the image, respectively. In terms of the scale variation of the UAV images, we use the acquisition of samples by UAVs at different altitudes in real scenes. On the uncertainty of UAV flight direction, a rotate augmentation method consistent with realistic UAV sampling is proposed to expand drone-view images.

The rotation augmentation method proposed in this paper does not rotate the image directly in the original image, which does not change the information contained in the image and does not match the actual UAV rotation scenario. As shown in the top panel of Figure 5. The target of the crop is the largest inline circle in the original image, and the largest inline square in that circle is cut out according to the rotation angle and then using a perspective transformation method to arrange the

picture horizontally. θ refers to the angle of rotation. This data augmentation not only expands the data in the direction of the drone's flight, but also enhances the generalisation of the model by erasing the image information at the edges, which is akin to applying some patches to the image, allowing the model to dig for more generic features. Besides, erasing the edge information allows the model to focus more on the information in the centre of the image, which is what we would expect as the centre point is closer to the real position of the drone. Six randomly rotated cropped images are shown on the bottom two rows of Figure 5. To verify the effectiveness of our proposed rotation augmentation method for UAV localization, we developed experiments for validation, as shown in Table 3, after the rotational change, the SDM metrics rise significantly by points.

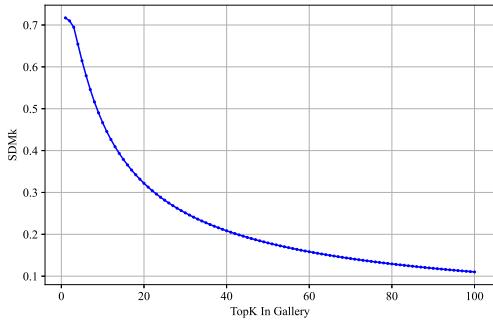


Fig. 6 Graph of the effect of K on the SDM indicator

Table 4 Table of test results for different dataset training weights in DenseUAV. FeatureDim refers to the dimensionality of the output feature vector.

Dataset	FeatureDim	SDM1	SDM3	SDM5	SDM10
ImageNet	2048	0.023	0.022	0.020	0.018
University-1652	512	0.046	0.042	0.038	0.033
DenseUAV	512	0.717	0.695	0.615	0.467

4.3 Impact of K in SDM

K in SDM_K refers to the K samples that are closest to the feature distance of query, and SDM_K

measures the spatial distance by these TopK samples and ignore the rest. Thus K determines the number of candidate gallery images that participate in the SDM calculation. To verify the effect of K on the SDM evaluation metric, we vary K from 1 to 100 to observe the trend of SDM. As shown in Figure 6, the curve decreases rapidly when $K > 3$, which is due to the fact that each query has only 3 corresponding gallery images in our dataset. in addition, SDM_K shows a monotonically decreasing trend.

4.4 Non-Dense Features vs. Dense Features

We tested features from different types of datasets in the DenseUAV test set, divided into Non-Dense and Dense types. (1) We used the ViTS model pre-trained in ImageNet as the standard model, with an output dimension of 2048. In addition, we trained the ViTS model in the University-1652 dataset, with an output dimension of 512. (2) We also use the ViTS model in the DenseUAV dataset, and the dimension of the output feature vector is also 512. The models were evaluated in the test set of DenseUAV using SDM evaluation metrics. As shown in Table 4, the results indicate that the features trained in ImageNet and University-1652 lack the ability to discriminate spatial locations, which is also related to the fact that the dataset itself is not targeted for dense tasks, which justifies the need for dense datasets to be proposed.

Table 5 SDM accuracy for different backbone training in DenseUAV and inference schedules for the testing phase.

Backbone	Inference Time	SDM1	SDM3	SDM5	SDM10
ResNet-50	1x	0.143	0.125	0.115	0.100
ResNet-101	1.25x	0.212	0.194	0.179	0.153
ResNeSt-50	1.2x	0.151	0.135	0.124	0.108
ViT-S	1.27x	0.454	0.436	0.390	0.305

4.5 Effect Of Backbone

To verify the impact of the backbone on the UAV localization task, we trained some backbone i.e. ResNet-50, ResNet-101, ResNeSt-50 [Zhang et al, 2020] and ViTS on the DenseUAV dataset and

tested them with SDM metrics. We also evaluate the time consumption of different backbone networks during the testing phase. All the backbone networks import the pre-training weights of ImageNet and only apply CrossEntropy Loss for training. As shown in Table 5, we found that the Transformer structure-based model had a stronger performance in the UAV localization task compared to the CNN structure. In terms of inference time cost, VitS has a test time of $1.27\times$, which is about the same as ResNet-101's $1.25\times$, but improves nearly 24 points in the SDM1 evaluation metric.

4.6 Compare With State-of-the-art Models

We adopted some existing state-of-the-art methods for training and testing in the DenseUAV dataset, and R@1, R@Top1%, mAP and our proposed SDM evaluation metrics were applied to evaluate the performance. As shown in Table 6, our Baseline improved nearly 29 points in R@1 compared to MSBA, nearly 12 points in R@Top1%, nearly 31 points in mAP, and nearly 29 points in our SDM1 metric. The experiments show that our baseline has a strong ability to identify spatial location information.

4.7 Robustness of flight altitude

Considering the height uncertainty of UAVs during actual flight, we designed experiments to verify the robustness of our baseline to UAV image scale variations. We divide the drone-view image sets of 80, 90, and 100m (relative ground height) from the test set and verify the results on the satellite gallery of DenseUAV. As shown in Table 7, the SDM metrics are close for different altitudes. The experiments show that our model has excellent robustness to the scale changes generated by the UAV flight altitude.

5 Application Analysis

Another contribution of this paper is to propose a UAV-based visual localization scheme and an improved neighbor-based retrieval method. Besides, this paper analyzes the reasons of why large distance deviations occur. In this paper, we introduce the general UAV localization scheme in

Section 5.1. And analyze the causes of mislocation in Section 5.2. And introduce the neighbor-serach UAV localization scheme in Section 5.3.

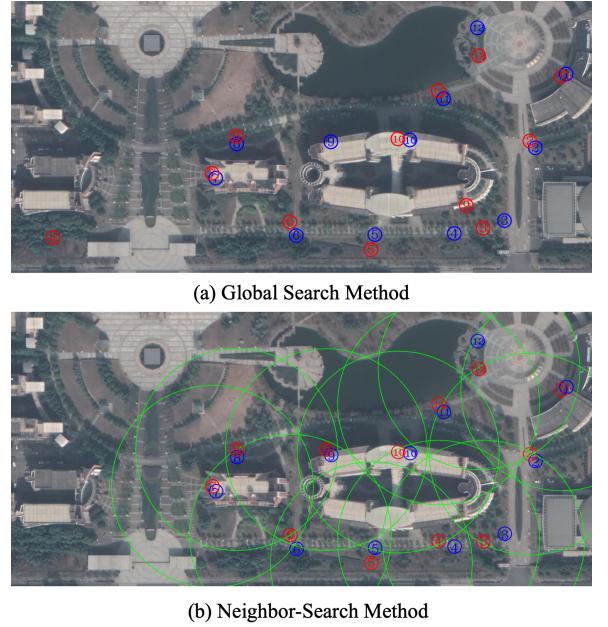


Fig. 7 Comparison of Global Search and Neighbor-Search methods.

5.1 Basic Global Search Method

The whole task of cross-view geo-localization is understood as an end-to-end system, with an input of a drone-view image and an output of the corresponding GPS information. For Visualization, we intercept part of the satellite map. As shown in Figure 7, the blur serial number represents the real GPS location of the UAV, while the red serial number is the GPS location predicted by the end-to-end system mentioned above, and the same number refers to a corresponding relationship. The global search method is subject to a large offset of the serial number 4 in Figure 7(a), for which we do a specific analysis in Section 5.2 for this kind of mislocation.

Most of the samples are predicted to be in the vicinity of the true position, which is because the displacement of the UAV in the air corresponds to a small change in its captured images. In addition, the satellite image library retrieved as a gallery is also discrete segmented. Our satellite images are sampled from google earth, and our sampling

Table 6 Results of different state-of-the-art models tested in DenseUAV with R@1, R@Top1%, mAP and SDM1 evaluation metrics.

Method	Feature Dim	R@1	R@Top1%	mAP	SDM1
University-baseline [Zheng et al, 2020a]	512	7.51%	56.2%	7.57%	0.155
Wehgited Soft Margin Triplet Loss [Liu and Li, 2019b]	512	9.70%	60.53%	9.72%	0.168
LPN [Wang et al, 2021]	2048	35.26%	85.33%	31.03%	0.415
MSBA [Zhuang et al, 2021]	3072	38.65%	86.79%	36.82%	0.430
Our Baseline	512	67.27%	98.54%	68.21%	0.717

Table 7 Effect of flight altitude on SDM indicators.

Height	SDM1	SDM3	SDM5	SDM10
80m	0.706	0.681	0.597	0.445
90m	0.724	0.706	0.623	0.471
100m	0.722	0.699	0.626	0.485

span is 1/4 of the satellite image size, for example, if the size of satellite image that we cropped is 640×640 , then the span between adjacent satellite images is 160 pixels. This is one of the most important reasons for the position shift in the matched images.

5.2 Error Localization Analysis

In the process of UAV localization, positioning deviations is caused by many objective reasons, we have mainly listed the following three points. First, there are many shadow areas in the satellite maps, which makes the fine-grained information of the images lost and increases the difficulty of geographic location matching. In addition, satellite images cannot be acquired in real time, and only historical satellite images can be obtained. The difference of time will lead to more or less changes between satellite maps and actual drone-view images, such as newly planted trees, newly built houses, and even changes in building color, which will lead to the non-correspondence of the contents in the images. Second, satellite images from different locations may also be very similar, and it can be understood to have a wrong match when the images contain very similar contents. To identify the cause of the mislocation we further studied point 4 in Figure 7(a), where a

large deviation appears. As shown in Figure 8, the red circles in the diagram refer to the position of the UAV and the blue circles with serial numbers refer to the position of the satellite images in ascending order of feature distance. The features of the mislocation satellite images have some similar characteristics to the query in the bottom left corner of the Figure 8. This situation will worsen as the satellite database grows larger. In view of this, we propose a localization strategy based on neighbor-search to solve the localization problem for large distance deviations. Third, the different flight heights of UAVs will lead to inconsistent image scales, for example, UAVs flying at 100 meters will contain more but smaller scenes than those flying at 80 meters, and this scale inconsistency will directly affect the accuracy of matching under extreme conditions.

5.3 Neighborhood Search Solution

The global search method not only generates positioning errors over large distances, but also increases the time consumption of post-processing. Relying on the spatial continuity of drone flight, we designed a neighbor-search method to improve this situation. The neighbor-search method is to narrow down the original global domain to a neighborhood domain. As shown in Figure 8, the yellow circle represents the location of the drone's last match and the green dotted line represents the area searched. It can be seen that the best match, satellite map number 1, was excluded from the search, while satellite maps 11, 12, 13, 17 and 20, which were not ranked highly before, became the new best choices. In terms of implementation, we reconstruct the candidate database by filtering

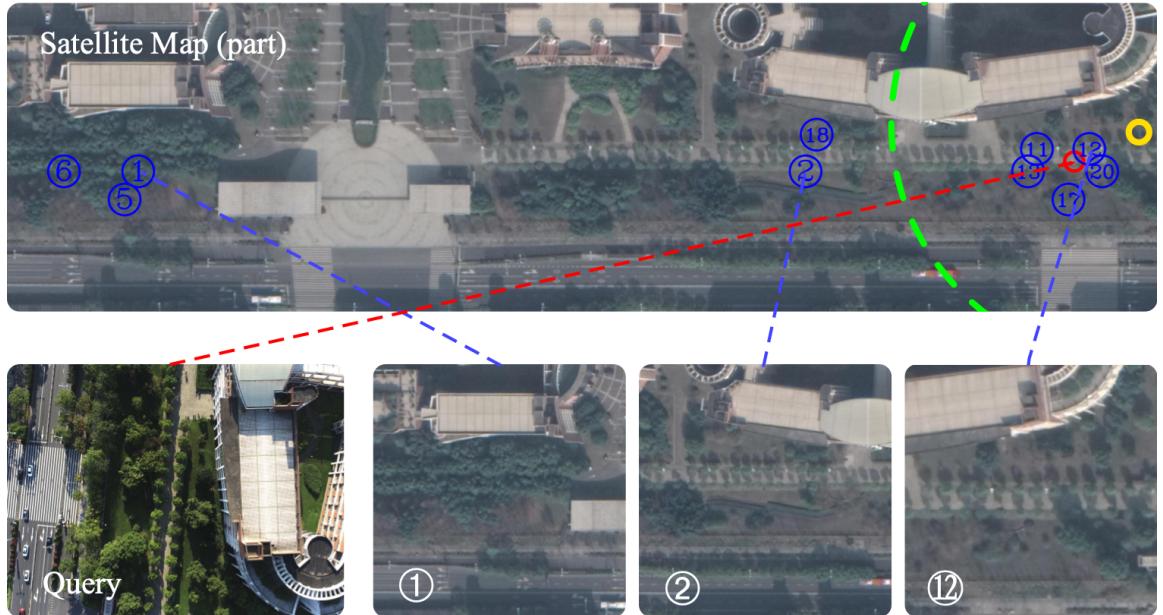


Fig. 8 Analysis diagram of the mislocation. The yellow circle is the position of the previous match, the green dashed line is the centre of the yellow circle with M as the search radius arc, the red circle is the real position of the current UAV, the blue circle is the position of the satellite image matched out of the current UAV image, and the serial number represents the ranking of the feature distance.

out satellite images whose spatial distance from the domain centroid is less than a certain threshold on the basis of the global satellite database. We then perform a search in this newly composed database for the satellite image with the closest feature distance. To clearly demonstrate the entire neighbor-search process, we visualise it in Figure 7(b). By comparing the results of the matching in Figure 7(a) and (b), we can conclude that the reduction of the retrieval domain will reduce the localization deviation. Reasonable selection of the area to be retrieved not only reduces the calculation time of the feature distance, but also improves the accuracy of the UAV positioning. However, this method of neighbor-search also has some limitations and will no longer be applicable when the spatial span of UAV sampling is large.

6 Conclusion

This paper contributes a dense UAV-based geolocation dataset named DenseUAV to solve the problem of UAV self-localization. Based on this dataset, a new evaluation metric for measuring spatial distance named SDM is proposed to evaluate the performance of the algorithm in

practical applications. In addition, a Transformer-based baseline is proposed to implement end-to-end training. In particular, a neiborgh-search post-processing strategy, which not only reduces the time consumption of UAV localization but avoids large position deviations, is proposed to optimize the UAV self-localization task.

Based on the idea of narrowing the retrieval range, there are many other schemes worth trying. The retrieval range can be further reduced according to the UAV's IMU and flight speed, thus further improving the positioning accuracy and reducing the positioning time consumption. In addition, the navigation problem will be solved after achieving the precise positioning of UAVs.

References

- Arandjelovic R, Gronat P, Torii A, et al (2016) Netvlad: Cnn architecture for weakly supervised place recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5297–5307
- Carion N, Massa F, Synnaeve G, et al (2020) End-to-end object detection with transformers. In: European Conference on Computer Vision, Springer, pp 213–229

- Dosovitskiy A, Beyer L, Kolesnikov A, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:201011929
- He S, Luo H, Wang P, et al (2021) Transreid: Transformer-based object re-identification. arXiv preprint arXiv:210204378
- Hu S, Lee GH (2020) Image-based geo-localization using satellite imagery. International Journal of Computer Vision 128(5):1205–1219. <https://doi.org/10.1007/s11263-019-01186-0>
- Hu S, Feng M, Nguyen RM, et al (2018) Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 7258–7267
- Jiang Y, Chang S, Wang Z (2021) Transgan: Two transformers can make one strong gan. arXiv preprint arXiv:210207074 1(3)
- Lin TY, Cui Y, Belongie S, et al (2015) Learning deep representations for ground-to-aerial geolocation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 5007–5015
- Liu H, Feng J, Qi M, et al (2017) End-to-end comparative attention networks for person re-identification. IEEE Transactions on Image Processing 26(7):3492–3506
- Liu L, Li H (2019a) Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5624–5633
- Liu L, Li H (2019b) Lending orientation to neural networks for cross-view geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5624–5633
- Liu Z, Lin Y, Cao Y, et al (2021) Swin transformer: Hierarchical vision transformer using shifted windows. arXiv preprint arXiv:210314030
- Long Y, Gong Y, Xiao Z, et al (2017) Accurate object localization in remote sensing images based on convolutional neural networks. IEEE Transactions on Geoscience and Remote Sensing 55(5):2486–2498. <https://doi.org/10.1109/TGRS.2016.2645610>
- Ma J, Jiang X, Fan A, et al (2021) Image matching from handcrafted to deep features: A survey. International Journal of Computer Vision 129(1):23–79. <https://doi.org/10.1007/s11263-020-01359-2>
- McManus C, Churchill W, Maddern W, et al (2014) Shady dealings: Robust, long-term visual localisation using illumination invariance. In: 2014 IEEE international conference on robotics and automation (ICRA), IEEE, pp 901–906
- Mehta S, Rastegari M (2021) Mobilevit: Light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:211002178
- Middleberg S, Sattler T, Untzelmann O, et al (2014) Scalable 6-dof localization on mobile devices. In: European conference on computer vision, Springer, pp 268–283
- Shi Y, Yu X, Campbell D, et al (2020a) Where am i looking at? joint location and orientation estimation by cross-view matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 4064–4072
- Shi Y, Yu X, Liu L, et al (2020b) Optimal feature transport for cross-view image geo-localization. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp 11,990–11,997
- Sun Y, Cheng C, Zhang Y, et al (2020) Circle loss: A unified perspective of pair similarity optimization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6398–6407
- Tian Y, Chen C, Shah M (2017) Cross-view image matching for geo-localization in urban environments. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 3608–3616
- Toker A, Zhou Q, Maximov M, et al (2021) Coming down to earth: Satellite-to-street view synthesis for geo-localization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6488–6497
- Touvron H, Cord M, Douze M, et al (2021) Training data-efficient image transformers & distillation through attention. In: International Conference on Machine Learning, PMLR, pp 10,347–10,357
- Vaswani A, Shazeer N, Parmar N, et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
- Vo NN, Hays J (2016) Localizing and orienting street views using overhead imagery. In: European conference on computer vision, Springer, pp 494–509

- Wang T, Zheng Z, Yan C, et al (2021) Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology* <https://doi.org/10.1109/TCSVT.2021.3061265>
- Wang X, Girshick R, Gupta A, et al (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
- Workman S, Souvenir R, Jacobs N (2015) Wide-area image geolocation with aerial reference imagery. In: Proceedings of the IEEE International Conference on Computer Vision, pp 3961–3969
- Yang F, Yang H, Fu J, et al (2020) Learning texture transformer network for image super-resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 5791–5800
- Yu H, Li G, Zhang W, et al (2020) The unmanned aerial vehicle benchmark: Object detection, tracking and baseline. *International Journal of Computer Vision* 128(5):1141–1159. <https://doi.org/10.1007/s11263-019-01266-1>
- Zhai M, Bessinger Z, Workman S, et al (2017a) Predicting ground-level scene layout from aerial imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 867–875
- Zhai M, Bessinger Z, Workman S, et al (2017b) Predicting ground-level scene layout from aerial imagery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp 867–875
- Zhang H, Wu C, Zhang Z, et al (2020) Resnest: Split-attention networks. arXiv preprint arXiv:200408955
- Zhao R, Ouyang W, Wang X (2013) Person re-identification by salience matching. In: Proceedings of the IEEE international conference on computer vision, pp 2528–2535
- Zheng L, Shen L, Tian L, et al (2015) Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision, pp 1116–1124
- Zheng S, Lu J, Zhao H, et al (2021) Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 6881–6890
- Zheng Z, Wei Y, Yang Y (2020a) University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In: Proceedings of the 28th ACM international conference on Multimedia, pp 1395–1403
- Zheng Z, Zheng L, Garrett M, et al (2020b) Dual-path convolutional image-text embeddings with instance loss. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)* 16(2):1–23
- Zhuang J, Dai M, Chen X, et al (2021) A faster and more effective cross-view matching method of uav and satellite images for uav geolocalization. *Remote Sensing* 13(19):3979. <https://doi.org/10.3390/rs13193979>