SUES-200: A Multi-height Multi-scene Cross-view Image Benchmark Across Drone and Satellite

Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang, Wenbo Hu

***Abstract*—The purpose of cross-view image matching is to match images acquired from the different platforms of the same target scene. With the rapid development of drone technology, how to help drone positioning or navigation through cross-view matching technology has become a challenging research topic. However, due to the existing public datasets don't include the differences in images obtained by drones at different heights, and the types of scenes are relatively homogeneous, which makes the models unable to adapt to complex and changing scenes. There is still potential to improve the accuracy of the models. We present a new cross-view dataset, SUES-200, to address these issues. SUES-200 contains images acquired by the drone at four flight heights and the corresponding satellite view images under the same target scene. To our knowledge, SUES-200 is the first dataset that considers the differences generated by aerial photography of drone at different flight heights. In addition, we build a pipeline for efficient training testing and evaluation of cross-view matching models. Then, we comprehensively analyze the performance of feature extractors with different CNN architectures on SUES-200 through an evaluation system for cross-view matching models and propose a robust baseline model on this dataset. The experimental results show that SUES-200 can help the model learn features with high discrimination at different heights.**

***Index Terms*—Cross-view Image Matching, Drone, Benchmark, Image Retrieval, Pipeline**

I. Introduction

CROSS-VIEW matching technique [1] is an essential research topic in computer vision, and this technique can be applied to many aspects, such as localization, navigation, autonomous driving, object detection. With more diverse ways of image acquisition, satellite and drone platforms play an important role in image acquisition. A standard cross-view matching work is as follows: given an image to be retrieved in the query dataset of one view, the matching system can find an image under the exact location in a large-scale candidate (gallery) dataset of another view. For cross-view matching under satellite and drone platforms, two main tasks need to be tackled: 1.Drone localization:Drone → Satellite. 2.Drone navigation:Satellite → Drone.The key of the cross-view matching technique is to learn the invariant and discriminative features of images under different views.

Most of the previous studies on cross-view matching [2]–[7] focus on the cross-view matching between street view and satellite view, or between street view and bird-view, for example, the datasets CVUSA [8] and CVACT [9] use panoramic street view and satellite view of the same target scene to form a cross-view image pair and build a deep neural network model to solve the problem of feature extraction. However, the quality of matching between street view and satellite image is usually limited to the small spatial scale of street view, easy to be obscured, and interfered with, which leads to the model not extracting the appropriate features.

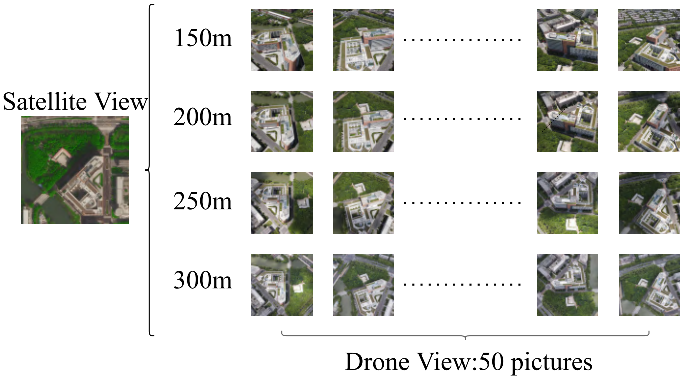


Fig. 1. A representative target scene of SUES-200 contains fifty drone view images from four heights and one satellite view image.

With the continuous development of drone technology [10]–[12], the flexibility and stability of drones are continuously improved, and using drone platforms can well describe the target scenes at different spatial and temporal scales. The traditional Drone view and satellite view image matching technology [13]–[15] is concentrated in the military field, where fixed-wing drones fly at higher flight height and collect images in real-time. The matching system match drone view image with satellite view image to infer the location of the drone. This autonomous positioning system is not affected by the external environment and has strong robustness in complex electromagnetic environments. However, such systems generally perform matching by extracting the hand-craft features [16]–[18] of images. Such feature extraction algorithms are less robust and susceptible to unfavorable factors such as lighting and occlusion, especially when drones fly at lower heights. There are many false matches or missing matches due to the excessive difference between the acquired images and the satellite view images.

Recently, new progress has been made in cross-view view matching research. Zheng et al. [19] establish the first drone-based multi-source cross-view matching dataset, University-

Runzhe Zhu, Ling Yin, Mingze Yang, Fei Wu, Yuncheng Yang are with the School of Electronic and Electrical Engineer, Shanghai University of Engineering Science, Shanghai 201602, China (email: m025120503@sues.edu.cn; lyin@sues.edu.cn; ymz871500142@163.com; fei_wu1@163.com; shawn.yangyc@foxmail.com.) Corresponding author: Fei Wu.

Wenbo Hu is with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (email: wenbohu@shu.edu.cn.)

1652, which contains three views: street view, drone, and satellite. It also publishes a baseline by designing a multi-branch CNN network. [20]–[24] conduct a more in-depth study of University-1652 and significantly improve the accuracy of the matching. However, University-1652 still has the following problems: 1.University-1652 uses synthetic images of drone views, which lack real-world lighting variations. 2.no attention is paid to the differences in drone images at different flight heights. 3.the captured scenes are of a single type, most buildings on the campus. These problems lead to poor differentiation of the features extracted by the model and poor robustness of the drone when flying at different heights.

Therefore, we propose a multi-height, multi-scene dataset across drone and satellite views based on the University-1652 dataset, called SUES-200. SUES-200 has the following features: 1.SUES-200 collects data from the same scene at four heights(150m, 200m, 250m, 300m). 2.SUES-200 contains more scenes, such as parks, schools, lakes, and public buildings. 3.SUES-200's drone view images are all captured in the real world, which is closer to the actual application. SUES-200 contains 200 target scenes, 120 scenes in the training set, and 80 scenes in the test set. Some of the data in SUES-200 are shown in Figure 1.

The traditional evaluation metrics for cross-view matching datasets are Recall@K [25] and AP, which cannot fit the new SUES-200 characteristics. Therefore, we design a new evaluation system that focuses on three aspects of model: 1.robustness at different heights. 2.preference for two tasks. 3.real-time performance during model inference. To address the problem that the training and testing process of previous cross-view matching models is complicated, and previous models are not appropriate to modify the model structure. We build a pipeline dedicated to cross-view matching, which helps us efficiently train, test, and evaluate models. In the experiment, we train and test feature extractors of different CNN architectures on SUES-200 through a pipeline. The model with the best overall evaluation results is released as the baseline model of SUES-200. We also evaluate the effect of multi-angle feature fusion on matching results and compare the performance of transfer learning models. Finally, the baseline model is extended and tested in the ablation experiment section. Our results show that SUES-200 can help the model learn high-level features in various and different heights scenes. As the flight height rises, the drone gradually becomes less affected by the environment and camera pose, and the performance metrics increase.

In Summary, the main contributions of this paper are as follows:

- We build a new cross-view matching dataset: SUES-200. The main feature of SUES-200 is that it contains different heights in the same target scene. All images are acquired in real environments of multiple types of scenes, including real-world light and shadow transformations and disturbances, which are closer to the actual application scenarios.
- We propose a new evaluation system based on the characteristics of SUES-200, which evaluates the robustness of the model at different heights, the preference for different tasks, and the real-time performance, in addition to the classical Recall@K and AP.
- We establish an efficient pipeline to train and test different CNN-based mainstream models and release the baseline model of SUES-200 according to the comprehensive evaluation results.

II. Related Work

Many previous cross-view datasets have focused on collecting images at the same location from different viewpoints via different platforms (e.g., panoramic cameras, satellites, drones, smartphones) to form image pairs and build datasets based on these image pairs. The dataset [2] utilizes publicly available data to create a cross-view dataset in which view one is the aerial view or "bird" view and view two is the street view, which contains a total of 78K data pairs. Tian et al. [4] mainly collected some locations in the city and constructed image pairs using "brid" view and the street view. In particular, this dataset also incorporates semantic information to label the corresponding buildings in the images from different views. They consider the differentiation of buildings a critical task in urban localization, so the module for object detection is accounted for in their network structure. In order to evaluate the model, they use PR curves and AP to evaluate the experimental results. CVUSA [26] a standard cross-view dataset consisting of panoramic street view and satellite view image pairs. Many previous works have been conducted based on this dataset. CVACT [9] is a larger panoramic dataset than CVUSA, with improved satellite image resolution and number of test sets compared to CVUSA, and with GPS-tag added to the corresponding scenes. Both CVACT and CVUSA use Recall@K to evaluate their matching results. In the field of multi-source cross-view scene matching, Zheng et al. [19] propose University-1652, the first geo-localization dataset based on the drone, which contains image data pairs consisting of satellite view-drone view-street view of 1652 buildings in 72 universities. University-1652 generally has one satellite viewpoint image, fifty-four drone viewpoint images, and multiple street view images at a location. Due to the unaffordable cost of the real-world flight, the drone viewpoint data in this dataset is obtained by simulated flight in Google Earth, where the drone simulation flight route is fly around the target scene and gradually drops in height. University-1652 adopts Recall@K and AP to evaluate their matching results. Inspired by the idea of University-1652, we collect the SUES-200 dataset, which emphasizes the differences in images acquired by drones at different heights and extends the types of scenes, all of which were acquired in real scenarios.

With the publication of the University-1652 dataset, progress has been made in the past year in cross-view matching algorithms based on drone views and satellite views. Liu el at. [20] propose LCM utilizes ResNet [27] as the backbone network and trains the image retrieval problem as a classification problem, and uses data augmentation to extend the satellite view images. The results show that LCM's Recall@1 and AP improve by 5-10 % over the baseline of University-1652. Wang el at. [22] design LPN after considering the contextual information of neighboring regions, which

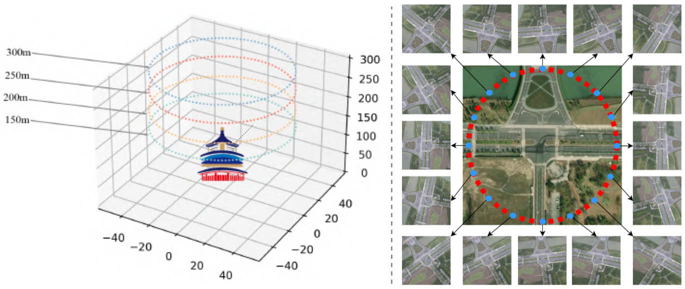


Fig. 2. The flight height of the drone when collecting images is 150m, 200m, 250m, 300m. The flight trajectory is one circle around the target scene

apply the square-ring partition strategy to divide feature maps. This strategy has good robustness to rotation changes. LPN achieves good performance on University-1652, CVACT, and CVUSA.Tian el at. [23] present a method that integrates the spatial correspondence between the satellite view and the surrounding area information, which consists of two parts, 1. converting the tilted view of the drone into a vertical view by perspective transformation. 2. making the image of the drone view closer to the satellite view by conditional GAN [28], the experimental results show that the method improves the accuracy by 5% over LPN on University-1652. Inspired by the attention mechanism, Zhuang el at. [21] develop MSBA in order to eliminate the differences in images acquired from different viewpoints. MSBA cuts the image into several parts with different scales, where the self-attention mechanism makes feature extraction more effective. The results show that MSBA performs better than LPN in accuracy and inference efficiency. We train cross-view matching models in the form of training classification models and then test and evaluate the performance of different backbone networks such as VGG [29], ResNet, DenseNet [30] in extracting features at different heights by pipeline.

III. SUES-200 Dataset

A. Dataset description

The cross-view matching dataset has the characteristics of multiple sources, multiple scenes, and panoramic views. We collect multi-source images of satellite views and corresponding drone views at 200 locations in the vicinity of our school. Specifically, to enable the model to learn highly discriminative features at different heights, we collect drone view images at 150m, 200m, 250m, and 300m. The rich multi-type scenes enable the models trained by the dataset to be applied in real environments. Therefore, SUES-200 selects a broader range of scene types, not limited to campus buildings but parks, schools, lakes, and public buildings. Another practical problem is that drones have to fly continuously over the same area, so the close proximity of some of the target locations chosen for the SUES-200 means that these targets will be very similar, as shown in Figure 2. To help the subsequent matching system, the cross-view matching model needs to distinguish these minor differences and extract valid invariant features in the image. In addition, the images acquired by the drone during the flight are usually one side of the target scene. In order to help the model extract high-quality features on different sides, the drone flies one lap along with the target with an onboard camera. The image of the drone's viewpoint in this scene consists of 50 frames of images evenly extracted from the flight video. The flight trajectory of the drone and the sampling process are shown in Figure 3.

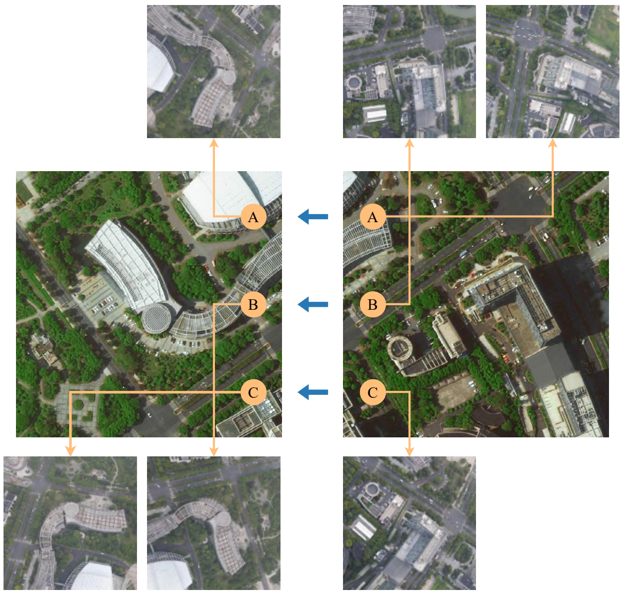


Fig. 3. A,B,C are the connection points of two consecutive satellite view images of the same area, and we give similar drone view images at the connection points.

In order to prevent information loss due to image resolution, both drone images and satellite images in SUES-200 use the original resolution of 1080 × 1080 and 512 × 512. There are 200 locations with 50 drone view images and 1 corresponding satellite view image for each location. SUES-200 is divided into training and test sets, where 60% is training data and 40% is test data. To accomplish the two tasks mentioned in the introduction, the test data include query drone dataset, query satellite dataset, gallery drone dataset, and gallery satellite dataset. Among them, the gallery dataset contains the test data and adds the training data as confusion data to increase the difficulty of matching. The comparison and statistics of the datasets are shown in Table I and Table II.

Finally, we summarize the new characteristics of the SUES-200 dataset:

1) **Multi-height:** SUES-200 contains data collected at different heights: 150m, 200m, 250m, 300m, and can evaluate model metrics at different heights. To our best knowledge, SUES-200 is the first cross-view dataset containing drone-view images from different heights.
2) **Multi-scene:** SUES-200 contains data from different types of scenes. It can help models extract invariant features in more scenes also expands the scope of scene applications for drone-based cross-view matching techniques.
3) **Continuous scenes:** Some of the SUES-200 target scenes are collected in the same area, so these scenes are continuous and similar, which is a challenge to the

TABLE I
COMPARISON BETWEEN SUES-200 AND OTHER CROSS-VIEW DATASETS.

Datesets	SUES-200	University-1652 [19]	CVUSA [26]	Tian et al [4].
Platform	Drone,Satellite	Drone,Ground,Satellite	Ground,Satellite	Ground,45° Aerial
Target	Diversity	Building	User	User
Height difference	TRUE	FALSE	FALSE	FALSE
Training	120 * 51	701 * 71.64	35.5k * 2	15.7k * 2
Images/Location	50 + 1	51 + 16.64 + 1	1 + 1	1 + 1
Evaluation	Recall@K & AP & Robustness & Preference & Real-Time	Recall@K & AP	Recall@K	PR&AP

TABLE II
STATISTICS OF SUES-200 TRAINING AND TEST SETS, INCLUDING THE IMAGE NUMBER AND THE SCENE NUMBER OF TRAINING SET, TESTING SET.

	Training Dataset		
Views	Locations	Images at Each Height	Total
Drone	120	6000	24000
Satellite	120	_	120
	Testing Dataset		
Views	Locations	Images at Each Height	Total
Drone query	80	4000	16000
Satellite query	80	_	80
Drone gallery	200	10000	40000
Satellite gallery	200	_	200

model's ability to differentiate, but it also more closely matches the actual drone flight environment.

B. Evaluation Protocol

This chapter introduces the evaluation system of SUES-200. In response to the existing real-world problems, in addition to the traditional Recall@K [3], [9], [31] and AP [4], [32] evaluation metrics, we propose 1.method to measure model robustness at different heights. 2.method to evaluate model preference for different tasks. 3.method to evaluate model real-time performance.

Recall@K and AP. SUES-200 contains a total of 200 target scenes, 120 scenes for training, and 80 scenes for testing. One hundred twenty scenes from the training are also included in the gallery as distractors. We note that there is no overlap between the training and testing data. Recall@K is very sensitive to the position of the first true-matched image appearing in the ranking of the matching result. Therefore, it is suitable for a test dataset that contains only one true-matched image in the candidate gallery. The AP is the area under the precision-recall (PR) curve, which considers the position of all true-matched images in the evaluation. The equations for Recall@K and AP are shown as follows:

$$\text{Recall@K} = \begin{cases} 1, & if\ order_{true} < K+1 \\ 0, & otherwise \end{cases} \tag{1}$$

$$\text{AP} = \frac{1}{m}\sum_{h=1}^{m}\frac{p_{h-1}+p_h}{2} \text{ where } p_0 = 1 \ p_h = \frac{T_h+1}{T_h+F_h} \tag{2}$$

Robustness. Since SUES-200 differentiates the images acquired by the drone at different heights, measuring the robustness of the model at different flight heights is also an important evaluation index. So we present a method to evaluate the robustness of the model based on different heights, which is calculated as follows.

$$\overline{y^i} = \frac{(y^i_{h_1}+y^i_{h_2}+y^i_{h_3}+y^i_{h_4})}{4} \tag{3}$$

y^i_h represents the recall@1 accuracy of model i at a certain height, $\overline{y^i}$ represents the average accuracy of model i at four heights.

$$c_i = \frac{1}{\sum\limits_{i=1}^{n=4}|y^i_h - \overline{y^i}|} \tag{4}$$

$$X = \{c_1, c_2, c_3, ..., c_m\} \tag{5}$$

For the set X we normalize all its elements between 0.4 and 0.6 and the result is indicated by X_{scaled}

$$X_{std} = \frac{X - min(X)}{max(X) - min(X)} \tag{6}$$

$$X_{scaled} = X_{std} \times (0.6 - 0.4) + 0.4 \tag{7}$$

Finally, the result of robustness of model i is expressed as:

$$\text{Robustness} = \frac{c_i \times (y^i_{h_1}+y^i_{h_2}+y^i_{h_3}+y^i_{h_4})}{4} \tag{8}$$

Preference. We consider that cross-view matching based on drone platforms and satellite platforms requires two tasks:Task1:Drone → Satellite,Task2:Satellite → Drone.There are significant differences in the performance of different models in these two tasks. Because the ratio of the drone to satellite images in the SUES-200 dataset is 50:1, the Recall@K accuracy of Task2 tends to be higher than that of Task1.

We propose the ”preference coefficient” as an indicator to evaluate the model's preference for two tasks. The closer the ”preference coefficient” is to 1, the more balanced the model performs in Task1 and Task2, and the larger it is, the stronger the model prefers Task2. The calculation formula is as follows:

$$\text{Adaption} = \frac{\sum\limits_{i=0}^{n=4}\frac{y^i_d}{y^i_s}}{4} \tag{9}$$

y^i_d represents the Recall@1 accuracy of Task 1, y^i_s represents the Recall@1 accuracy of Task 2, and $n = 4$ represents the task at four heights.

Real-time. In the actual application process, there will be requirements for the real-time performance of the cross-view

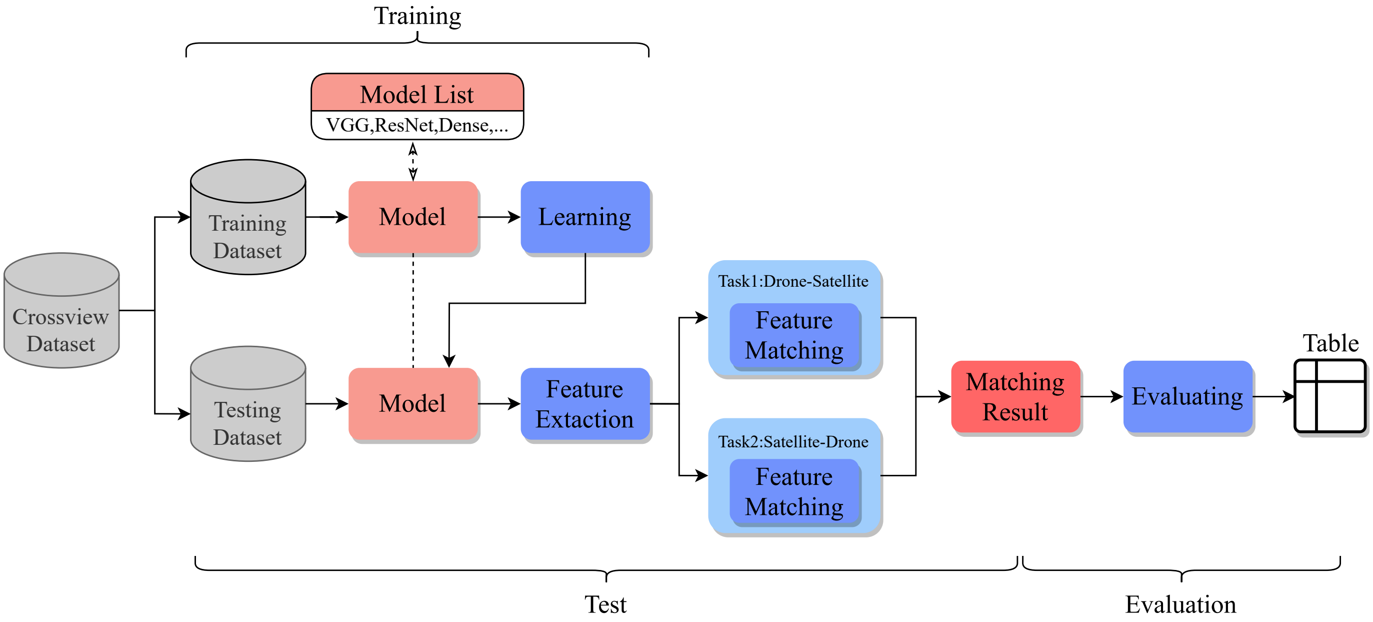


Fig. 4. Pipeline reads images from the dataset and sends them to the currently selected model for training. After training, the model with the best parameters is selected and sent to Task1 or Task2 for testing, and the evaluation module evaluates the test results to form an evaluation table.

matching model. Therefore, we refer to the idea of [21] to evaluate the real-time performance, choose a base model with better real-time performance, and take its inference time as the benchmark time which set to 1.00, and the other inference The time is then denoted as $1.00 \times T(0 < T < +\infty)$.

IV. Method

A. Pipeline

We build a Pipeline to solve the cross-view matching problem, which is used to train and test different models and build evaluation systems efficiently. As shown in Figure 4, in this Pipeline, the leftmost input is the cross-view matching dataset, and the rightmost output is the values of each evaluation index, in which the model is a deep neural network built by the user. The network is divided into a backbone network and a classification network part. Selecting different feature extractors in the “Model List” will replace the backbone network part in the corresponding network structure, and the user can also customize its network structure. The details of the network structure of deep neural networks will be illustrated in the next section. The images in the test set are input to the model to extract features and complete Task1:Drone → Satellite,Task2:Satellite → Drone, and the obtained feature matching results are finally passed to the evaluation unit to get the evaluation table.

B. Network architecture and loss function

The drone and satellite images included in SUES-200 originate from different data sources, but there are still some similarities. Our goal in designing the deep learning network is to extract robust and invariant features in both images separately and map them to a high-dimensional space to help the following matching process. After referring to previous studies, similarly, we build a two-branch deep convolutional neural network architecture, where one branch is used to extract features from satellite view images, and the other branch is used to extract features from drone view images. To test the effect of different CNN structures on different source image feature extractors, we refer to the CNN structures that extract features in two-branches as backbone networks, and these backbone networks are able to be replaced. In the training process, we add a fully connected layer with a softmax layer at the end of the branch to treat it as a classification task. Each target location is treated as a class to train the whole network. The network structure is shown in Figure 5.

In recent years, different CNN structures have been greatly developed. ResNet [27] is widely used as the backbone network [20]–[23] for feature extraction in the field of cross-view matching due to its clever design structure and excellent performance. With the further research on ResNet and the emergence of attention mechanism, some scholars have further improved ResNet, such as SE-ResNet [33],ResNeSt [34],CMAB-ResNet [35], and these models have achieved excellent performance on image classification datasets such as ImageNet [36]. In addition to ResNet, other structured CNNs are also a hot topic of research, e.g., DenseNet [30], EfficientNet [37], Inception [38]. Is there a more proper feature extractor than ResNet in cross-view matching? In our experiment, we test the improved ResNet and other CNN structures on SUES-200 and evaluate these models according to the evaluation system mentioned above.

For the loss function, because the model training process is considered to deal with a multi-classification task, we adopt cross-entropy as the loss function, which is typical in multi-classification tasks. Cross entropy is mainly used

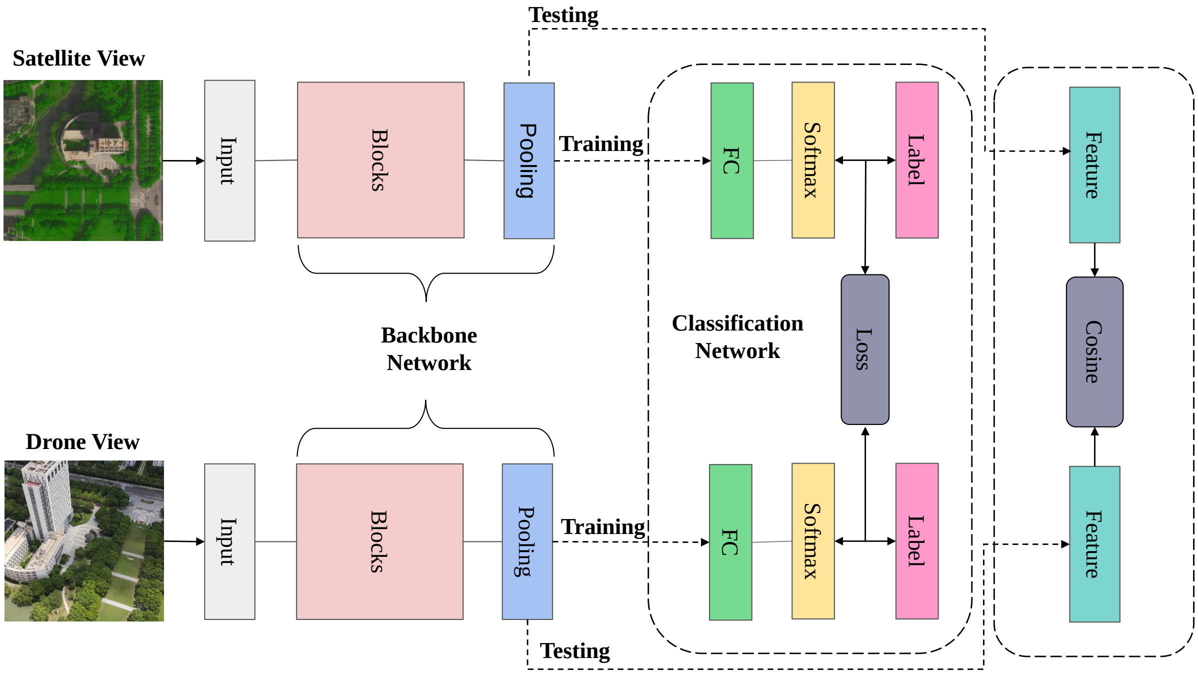


Fig. 5. The basic Network architectures for cross-view matching. We apply two-branch CNN structures with cross entropy loss to train model.The cosine distance is used to calculate the similarity between the query and candidate images in the gallery.

to determine how close the actual output is to the expected output, i.e., the smaller the cross entropy between the network output and the label, the better the classification ability of the network.$z_j^i(y)$ is the logarithm of ground-truth$y$, and $\hat{p}(y|x_j^i)$ is the the probability of predicted outcome of the model equal to ground-truth y. The mathematical formula is shown:

$$\hat{p}(y|x_j^i) = \frac{exp(z_j^i(y))}{\sum_{c=1}^{C} exp(z_j^i(c))} \tag{10}$$

$$\text{Loss} = \sum_{i,j} -log(\hat{p}(y|x_j^i)) \tag{11}$$

In the two-branch CNN, both outputs of the model need to be compared with the label and get two loss values. We let the loss of drone view be L_d, and the loss of satellite view be L_s, these two loss values are added to get L_{total}. We optimize the whole network through L_{total}. The equation expression is shown as follows:

$$L_{total} = L_s + L_d, \tag{12}$$

In the test stage, the query images in the test set are from drone view and satellite view. We feed the query images to the model with fixed parameters, remove the classification network from training layer, and make the backbone network output the feature vectors directly. The feature vector of drone view is represented as f_d, and the feature vector of satellite view is represented as f_s. Our test aims are to find the most similar set of feature vectors by cosine distance to measure the similarity between f_d and f_s. f_di and f_si are part of the feature vector, and a smaller cosine distance means that the set of features is less similar. The larger cosine distance means a smaller distance between the two features and a greater correlation between the corresponding features. The formula is shown as follows:

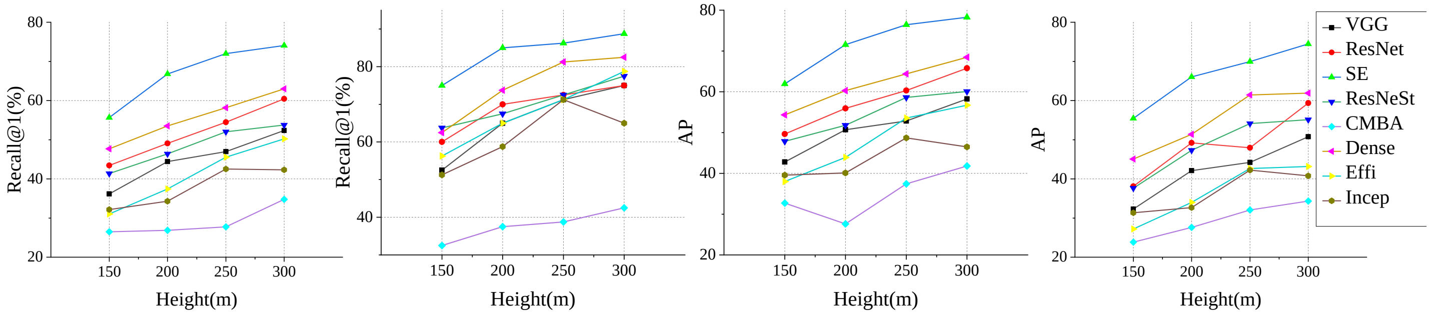
$$\text{Cosine} = \frac{f_d f_s}{||f_d|| \times ||f_s||} = \frac{\sum_{i=1}^{n} f_{di} f_{si}}{\sqrt{\sum_{i=1}^{n} (f_{di}^2)}\sqrt{\sum_{i=1}^{n} (f_{si}^2)}} \tag{13}$$

V. Experiment

This chapter first describes the experimental setting and details, followed by a comprehensive evaluation of multiple feature extractors through Pipeline. The impact of multi queries on the matching performance is explored. In addition, we test the performance of the transfer learning model on SUES-200. Finally, we reproduce some classical cross-view matching models on SUES-200.

A. Implement Details

Different feature extractors are used in our backbone network, and all of them are loaded with ImageNet's pre-trained weights to speed up the convergence of the model. However, the amount of work required to tune so many models to the optimum is incalculable, so for training, we applied the grid



(a) (b) (c) (d)

Fig. 6. The Recall@K accuracy curve and AP value curve at 150m, 200m, 250m, and 300m. (a): Recall@1 curve of Drone → Satellite. (b):Recall@1 curve of Satellite → Drone. (c):AP curve of Drone → Satellite. (d):AP curve of Satellite → Drone.

TABLE III
THE ROBUSTNESS OF DIFFERENT BACKBONE NETWORKS.

Backbone	Drone → Satellite Robustness	Satellite → Drone Robustness
VGG16-bn	21.74	26.37
ResNet-50	23.43	32.51
SE-ResNet-50	29.80	40.42
ResNeSt-50	23.77	32.95
CBAM-ResNet-50	17.38	22.69
DenseNet-201	26.23	30.44
EfficientNetv1-b4	17.40	27.13
Inceptionv4	18.43	25.38

search method to search for the following hyperparameters in the network: learning rate, drop out rate, weight decay. The image size is resized to (384,384) before feeding to the network, and only the basic image augmentation methods are used: Random Crop and Random Horizontal Flip. The optimizer of the neural network is SGD (momentum=0.9), and the initial learning rates of the backbone network and the classification network are set to 0.1 times and 1 times of the learning rate. The learning rate decay is MultiStepLR, and the parameters of the classification network are initialized with Kaiming Initialization [39]. Our model is built basing Pytorch, and all experiments are conducted on an NVIDIA RTX TiTAN GPU.

B. Evaluation of Different Extractors

Can SUES-200 help the model learn highly discriminative features? Can Pipeline efficiently perform the tasks from training, testing to evaluation? In this section, we set up experiments to comprehensively evaluate feature extractors of different CNN architectures and use the model with the best experimental results as the baseline model of SUES-200.

Recall and AP. With the help of Pipeline, we could quickly train on the SUES-200 to test and evaluate the models. As shown in Figure 6, we compare the feature extraction capability of different backbone networks by Recall@K and AP. SE-ResNet achieved the best performance at all four flight heights. In the drone view target localization task (Drone → Satellite), Compared with ResNet, the accuracy of Recall@1 increases from 43.42%, 49.42%, 54.47% 60.43% to 55.65%,66.78%,72%,74.05%, the value of AP raises from 49.65,55.91,60.31,65.78 to 61.92,71.55,76.43,78.26 in four heights. In in the drone navigation task (Satellite → Drone), Compared with ResNet, the accuracy of Recall@1 increases from 57.50%,68.75%,72.50%,75.00% to 75%,85%,86.25%,88.75%, the value of AP raises from 38.11,49.19,47.94,59.36 to 55.46,66.05,69.94,74.46 in four heights. The results show that as the flight height of the drone increases, the drone is less affected by environmental disturbances and its camera pose. The images camera acquires become more similar to satellite images, and the Recall@K and AP of the model improve. Compared with ResNet, SE-ResNet with SE module can greatly improve the feature extraction ability of the model's backbone network. Compared with other CNN models, such as other versions of ResNet: CBAM-ResNet, ResNeSt, or the new design ideas of EfficientNet and Inception, these networks have deeper and more complex network structures, and previous results have shown that they can achieve excellent performance on ImageNet. However, they do not achieve better results on SUES-200, probably because the features they extract were not suitable for cross-view matching tasks.

Robustness. As the flight height of the drone affects the accuracy of the matching system, in order to measure the robustness of the model to perform positioning or navigation tasks at different flight heights. We evaluate the robustness of the model by designing Equations (3)-(8). As can be seen from Table III, In Task1 (Drone → Satellite), seresnet achieves the highest value of 29.80, indicating that seresnet can complete the localization task with high accuracy and strong robustness at different heights. In Task2 (Satellite → Drone), the robustness index of seresnet is 40.42, indicating that seresnet is able to extract the required robust features at different heights in the navigation task.

Preference. We believe that it is essential for practical applications to choose a suitable model for Task1 or Task2. Therefore, the primary purpose of the ”preference coefficient” is to measure whether different models have a preference for Task1 and Task2. Therefore, the more balanced the performance of the model in Task1 and Task2, and the larger the ”preference coefficient” is, the stronger the model's preference for Task2 is. As shown in Figure 7, seresnet has a stronger adaptive ability for both Task1 and Task2.

TABLE IV
THE MATCHING ACCURACY (%) OF MULTIPLE QUERIES BASED ON THE BASELINE. 50,25,10,5,1 DENOTE MULTIPLE-QUERY IMAGE SETTING

	Drone → Satellite							
Query	150m		200m		250m		300m	
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP
50	66.25	71.61	77.50	80.84	80.00	83.84	82.50	85.49
25	64.38	69.98	76.25	79.87	79.37	83.22	80.00	83.54
10	62.00	67.89	74.25	78.18	77.00	80.77	79.00	82.58
5	61.13	66.84	72.75	76.73	76.25	80.06	77.50	81.29
1	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26

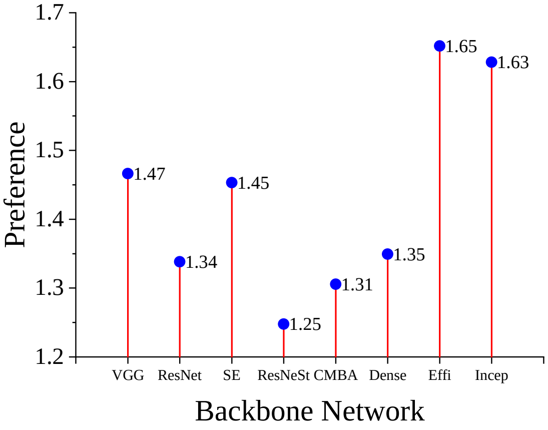


Fig. 7. The preference of different backbone networks. SE-ResNet strikes a balance between Task 1 and Task 2.

Real-time. In the model inference phase, real-time is a vital evaluation metric for the model, and it also directly determines whether the model can be put into practical application. Therefore, we evaluate the inference speed of different models under two tasks, it can be seen from Table V. We take the inference time of ResNet as the base time: 1.00. We can learn that VGG spends the most time on inference, and Task1 and Task2 are 1.18 and 1.17 times the base time, respectively, while the other models still obtain similar inference times with differences in the number of parameters.

TABLE V
THE NUMBER OF PARAMETERS OF ALL MODELS AND THEIR REAL-TIME WITH BENCHMARK

Backbone	Params(M)	Drone → Satellite	Satellite → Drone
VGG16-bn	272.86	1.18	1.17
ResNet-50	49.24	1.00	1.00
SE-ResNet-50	54.30	1.02	1.02
ResNeSt-50	53.09	1.02	1.00
CBAM-ResNet-50	59.30	1.04	1.02
DenseNet-201	35.73	1.05	1.02
EfficientNetv1-b4	37.06	1.01	1.00
Inceptionv4	83.98	1.03	1.01

C. Multiple Queries

Does multi-angle feature fusion improve the efficiency of matching? In previous matching experiments, a single drone-view image was used as a query for Drone → Satellite. Each scene in the SUES-200 dataset provides a full 360-degree view of the drone view image, which provides complete and comprehensive information about the target scene from different views. Therefore, if one query cannot describe the target scene, we can use multiply drone view images as queries at the same time to explore whether these multi-view query images can improve the accuracy and precision of matching. In the evaluation, we average the features obtained from multiple images and use the average features they obtain as the query. It can be seen from Table IV, we set the multiple-query image to 50, 25, 10, 5, 1, and the experimental results show that the multiply queries contain more and more images, and the Recall@K and AP of model matching are improved accordingly. When the average features of 50 images are used as queries, compared with the queries of single images from 150m to 300m, the accuracy is improved by 10.60%,10.72%,8%,8.45% . It also shows that the multi-angle features are more helpful for drone localization tasks when flying at lower heights.

D. Transfer Learning

Can previous datasets help the model learn features at different heights? Do pre-trained weights have an impact on the model training? We test whether the models obtained from training on the ImageNet dataset, as well as the University-1652 dataset, can extract discriminative features at different heights. As a control, we take training from scratch on SUES-200 with ImageNet as pre-trained weights and University-1652 as pre-trained weights. The backbone networks in the above networks are all SE-ResNet50. As can be seen from Table VI, the University-1652-based transfer learning model achieves surprising results compared to ImageNet, which validates that University-1652 can be applied to real scenes. But University-1652's ability at different heights is still limited because the dataset does not distinguish the effects of different heights. Further, we find that the model trained from scratch is much less capable of extracting features than the model trained based on ImageNet. Another interesting finding is that the model starting training based on the pre-trained weights of University-1652 perform better than the one based on ImageNet, which also shows that the initialization weights of the model are significant.

E. Other Baseline Model in SUES-200

How does the classical cross-view matching model perform on the SUES-200? Some previous works [20], [22]

TABLE VI
TEST RESULTS OF TRANSFER LEARNING MODELS AND PRE-TRAINED WEIGHTS ON SUES-200.

Drone → Satellite								
Traning Set	150m		200m		250m		300m	
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP
ImageNet	8.65	12.19	10.13	13.95	13.55	17.96	14.27	18.84
University-1652	32.33	39.01	40.55	47.25	45.63	52.27	50.05	56.26
SUES-200(from scratch)	7.40	12.44	8.90	14.03	8.05	13.94	7.90	14.22
SUES-200(ImageNet pre-trained)	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
SUES-200(U1652 pre-trained)	62.30	67.45	69.10	73.88	76.50	80.69	80.08	83.36

Satellite → Drone								
Traning Set	150m		200m		250m		300m	
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP
ImageNet	10.00	6.79	7.50	6.38	18.75	11.96	26.25	16.00
University-1652	25.00	23.69	37.50	32.49	43.75	39.19	48.75	41.81
SUES-200(from scratch)	11.25	8.01	11.25	9.55	10.00	10.21	12.50	9.85
SUES-200(ImageNet pre-trained)	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
SUES-200(U1652 pre-trained)	80.00	60.62	83.75	72.28	88.75	77.84	88.75	80.08

TABLE VII
TEST PERFORMANCES OF LCM AND LPN ON SUES-200

Drone → Satellite								
Methods	150m		200m		250m		300m	
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP
SUES-200 baseline	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26
LCM [20]	43.42	49.65	49.42	55.91	54.47	60.31	60.43	65.78
LPN(block=4) [22]	61.58	67.23	70.85	75.96	80.38	83.80	81.47	84.53

Satellite → Drone								
Methods	150m		200m		250m		300m	
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP
SUES-200 baseline	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46
LCM [20]	57.50	38.11	68.75	49.19	72.50	47.94	75.00	59.36
LPN(block=4) [22]	83.75	66.78	88.75	75.01	92.50	81.34	92.50	85.72

design deep neural networks that achieve excellent performance on different cross-view matching datasets. We mainly select two works in the cross drone view and satellite view domains, migrated their backbone network designs into our pipeline, and put our dataset into pipepline for training. The experimental results are shown in Table VII. Due to the feature partitioning strategy presented by the LPN for extracting semantic information, the strategy is able to extract global features of the image instead of focusing on the center of the image alone. LPN achieves excellent performance on both University-1652 and SUES-200, especially Task1, which has 6% - 8% improvement at each height.

VI. ABLATION STUDY

A. *Effect of image size*

Do different image resolutions cause loss of image information? The resolutions of the drone and satellite images in the SUES-200 dataset are 512×512 and 1080×1080, both of which contain much unused detail information. Therefore, in the ablation learning phase, we resize the input images to 512×512 and 256×256, keeping all other conditions constant, as shown in Table VIII. When we increase the resolution to 512×512, there is a 1%-2% improvement in both tasks at each height. Meanwhile, its real-time performance is 20% lower than before at 384×384, and there is no advantage in robustness. Moreover, when we reduce the resolution to 256×256, the performance of both tasks show a large decline. Indicating that a large amount of useful information is lost in the image at 256×256, which lead to the model's inability to extract valid features.

B. *Effect of sharing weights*

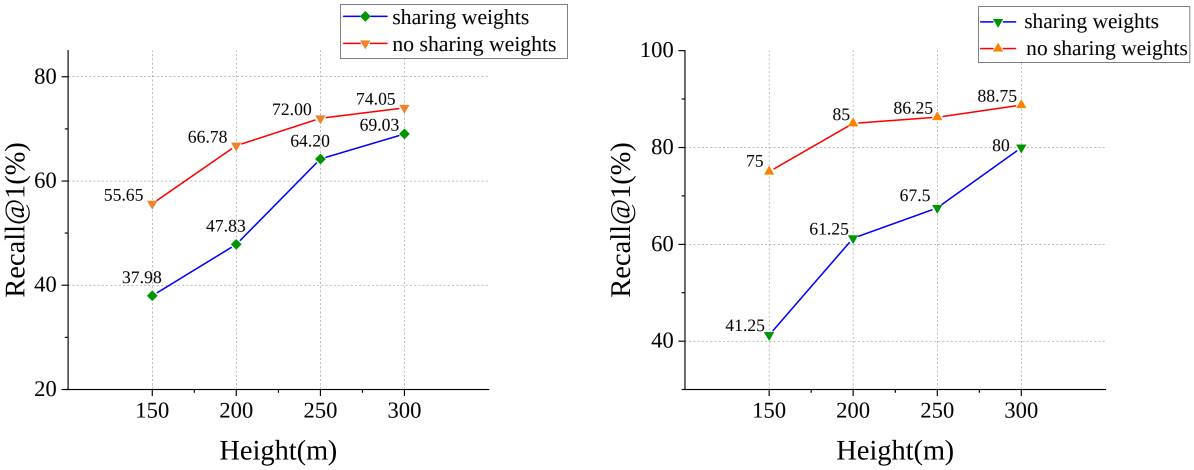
Do sharing weights help the model learn better features? As the flight height of the drone rises, the drone and satellite images will become more and more similar, so is it possible to improve the model learning efficiency by sharing the weights of both branches? We test the effect of sharing model weights on the final test results in the baseline model. Figure 8 show that the evaluation metrics of both tasks show significant decreases when the sharing weights are available, but the difference values between the sharing and unsharing weights decrease as the drone flight height increases. As the drone's height rises, images collected by the drone are more and more similar to satellite view images. A possible explanation for this might be that sharing weights can help model extract more efficient features in similar images.

C. *Effect of different loss function*

Do different loss functions affect the learning effect of the model? The most common loss functions in previous studies of matching retrieval tasks are contrastive loss [40] and triplet loss [41], and these loss functions achieve good performance in other works. To verify the feasibility of these

TABLE VIII
ABLATION STUDY ON IMAGE SIZE DURING INFERENCE ON SUES-200.

Drone → Satellite										
Image size	150m		200m		250m		300m		Time	Robustness
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP		
256	44.38	50.65	56.52	62.09	62.75	68.35	66.30	71.69	0.84	24.12
384	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26	1.02	29.80
512	54.85	61.04	65.25	70.72	74.23	78.67	79.18	82.68	1.30	27.35
Satellite → Drone										
Image size	150m		200m		250m		300m		Time	Robustness
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP		
256	63.75	37.85	80.00	54.48	82.50	62.71	82.50	63.91	0.85	31.57
384	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46	1.02	40.42
512	76.25	55.04	86.25	66.81	88.75	70.98	92.50	75.67	1.28	38.69



(a) (b)

Fig. 8. The accuracy of Recall@1 without sharing weights is always higher than that of Recall@1 with sharing weights, but the gap decreases as the height rises.(a) Drone → Satellite (b)Satellite → Drone

TABLE IX
ABLATION STUDY ON LOSS FUNCTION DURING INFERENCE ON SUES-200

Drone → Satellite									
Loss	150m		200m		250m		300m		Robustness
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP	
CrossEntropy [40]	55.65	61.92	66.78	71.55	72.00	76.43	74.05	78.26	29.80
Contrastive [41]	56.40	62.23	65.75	71.21	72.80	77.23	76.72	80.97	28.74
Triplet(margin=0.3)	57.25	62.92	63.27	68.92	71.07	75.94	73.83	78.00	29.15
Satellite → Drone									
Loss	150m		200m		250m		300m		Robustness
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP	
CrossEntropy [40]	75.00	55.46	85.00	66.05	86.25	69.94	88.75	74.46	40.42
Contrastive [41]	77.50	55.22	82.50	66.74	85.00	65.72	88.75	75.64	44.76
Triplet(margin=0.3)	75.00	52.13	82.50	60.54	87.50	70.03	88.75	74.73	39.10

loss functions on our baseline model, we strictly keep the backbone network as well as other parameters constant during the experiments. From Table IX, we observe that each of these three loss functions has its advantages and disadvantages in terms of Recall@K and AP. However, when evaluating the robustness metric, cross-entropy loss achieves the best score of 29.80 in Task1 and contrastive loss achieves 44.76 in Task2.

VII. Visualization

In this section, we visualize the retrieval results of SE-ResNet on SUES-200. Figure 9 shows the visualization results of the baseline model under Rank 5 at different heights and two tasks. It can be seen that the accuracy of model retrieval keeps improving with the rise of height and the ability to distinguish similar scenes increases. Furthermore, we also visualize the heat maps generated by different models on SUES-200. Figure 10 compare the results of ResNet, Dense, and SE-ResNet on drone view and satellite view. From left to right: the original image under drone view, the heat maps on different heights, the heat map under satellite-view. The heat maps show that Dense's activation area is larger than ResNet, while the area activated by SE-ResNet is more consistent with

TABLE X
ViT [42] achieves a outperform result on SUES-200.

Drone → Satellite									
Methods	150m		200m		250m		300m		Time
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP	
ViT-b-p16	59.32	64.94	62.30	67.22	71.35	75.48	77.17	80.67	2.46
Satellite → Drone									
Methods	150m		200m		250m		300m		Time
	Recall@1	AP	Recall@1	AP	Recall@1	AP	Recall@1	AP	
Vit-b-p16	82.50	58.88	87.50	62.48	90.00	69.91	96.25	84.10	2.48

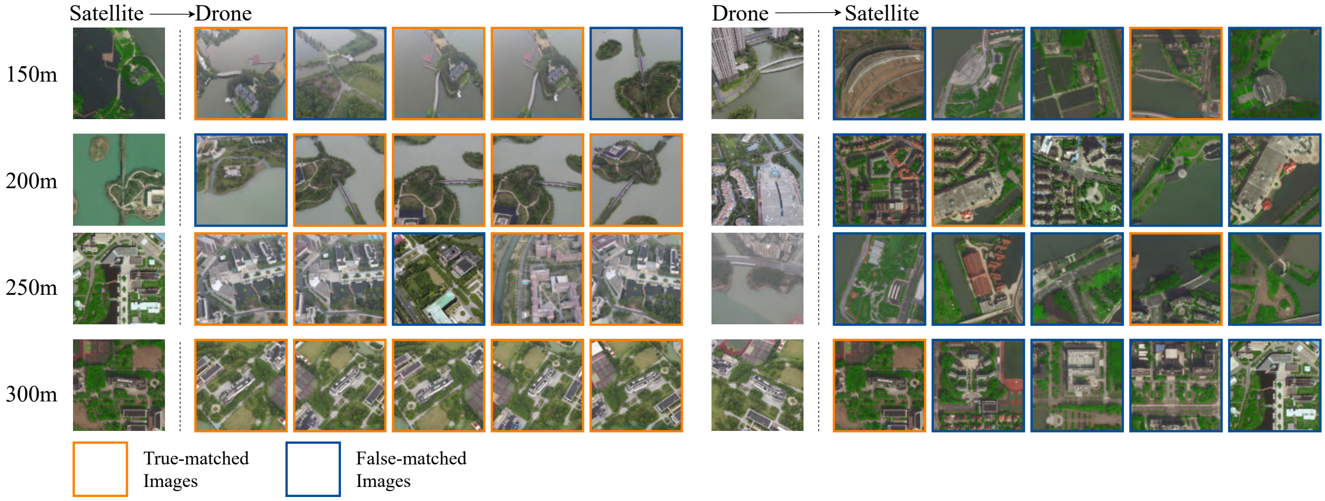


Fig. 9. Qualitative image retrieval results. Top-5 retrieval results of drone view target localization on SUES-200. Top-5 retrieval results of drone navigation on SUES-200.

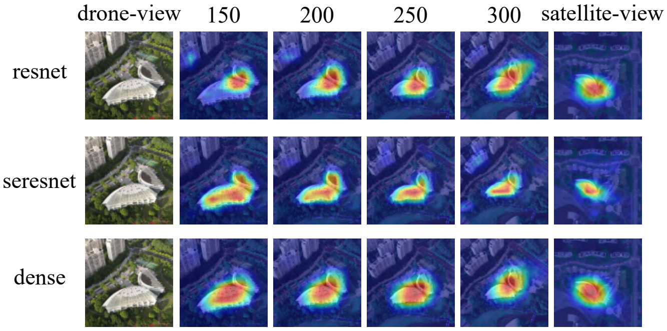


Fig. 10. Visualization of heatmaps. Heatmaps are generated by ResNet, baseline and Dense

the shape of the main target in the scene.

VIII. Discussion

In this study, we find significant differences in the cross-view matching results between the drone view and the satellite view at different heights. At the heights of 150m and 200m, the drone is more influenced by the surrounding environment and the camera pose, and the acquired images are very different from the satellite view images. So there is a low accuracy when the drone flight at a low height. However, as the flight height of the drone rises, the drone is less influenced by the environment and the camera pose, and the accuracy of feature matching gradually increases. At the same time, we believe that the bottleneck of previous research on cross-view matching studies lies in the lack of a suitable feature extractor. So we test the feature extraction effect of different feature extractors through our pipeline. The data show that the ResNet with the module of Squeeze-and-Excitation module can extract the most robust features with the best overall performance. SUES-200 distinguishes differences in drone flight at different heights. Our work fills a gap in cross-view matching field and includes a wider variety of scenarios than previous work.

However, SUES-200 still suffers from a small number of samples and a limited degree of flight height differentiation. Our baseline model directly classifies the feature maps from the backbone network, lacking consideration for the offset and image size changes that occur when aerial images of drones are taken. In the past few years, transformer architecture [43] has made a breakthrough in the field of vision [42], [44], [45], and we also try to use ViT [42] as the backbone network to extract features, as shown in Table X, and the experimental results greatly exceeded our expectations, which shows that transformer architecture has great potential in the field of cross-view matching. In the future, the main issues to be considered are how to solve the offset generated by the drone flight and how to adapt to the impact of different heights on the drone aerial images. We believe that there is a large amount of redundant information in the images captured by the drone, and that this redundant information may be caused by the offset of the drone during flight or by interference from the surrounding environment. Therefore, we hope to design a deep neural network to filter out the invalid redundant information in the images.

IX. Conclusion

In conclusion, Our study investigates the problem of image matching across drone and satellite views at different heights. We propose a multi-height, multi-scene benchmark called SUES-200, which contains multi-height drone and satellite view images for 200 locations. We also present new evaluation metrics and a pipeline to assess the effectiveness of the new model on the SUES-200. Our experiments find that the accuracy and precision of matching increase as the drone's flight height rises. In addition, after evaluating different feature extractors, we publish the model with the best overall performance as the baseline model of SUES-200. We also observe that appropriate pre-trained weights and multi-angle image fusion can help the model achieve even better result, pointing the way to improve matching efficiency further. In the future, we will continue our research on cross-view matching to eliminate the interference information in drone aerial images and improve the performance of drones at low heights.

References

[1] T. Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," *Springer International Publishing*, 2016.

[2] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5007–5015.

[3] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 867–875.

[4] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3608–3616.

[5] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[6] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 11 990–11 997.

[7] Y. Shi, X. Yu, D. Campbell, and H. Li, "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*, 2020, pp. 4064–4072.

[8] S. Workman and N. Jacobs, "On the location dependence of convolutional neural network features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 70–78.

[9] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 5624–5633.

[10] L. Wang, J. Li, B. Huang, J. Chen, X. Li, J. Wang, and T. Xu, "Auto-perceiving correlation filter for uav tracking," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[11] C. Zhan, H. Hu, X. Sui, Z. Liu, J. Wang, and H. Wang, "Joint resource allocation and 3d aerial trajectory design for video streaming in uav communication systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 8, pp. 3227–3241, 2020.

[12] Y. Sun, B. Cao, P. Zhu, and Q. Hu, "Drone-based rgb-infrared cross-modality vehicle detection via uncertainty-aware learning," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2022.

[13] C. ZHAO, Y. ZHOU, Z. LIN, J. HU, and Q. PAN, "Review of scene matching visual navigation for unmanned aerial vehicles," *SCIENTIA SINICA Informationis*, vol. 49, no. 5, pp. 507–519, 2019.

[14] X. Zhuo, T. Koch, F. Kurz, F. Fraundorfer, and P. Reinartz, "Automatic uav image geo-registration by matching uav images to georeferenced image data," *Remote Sensing*, vol. 9, no. 4, p. 376, 2017.

[15] D. L. Krishnan, K. Kannan, R. Muthaiah, and M. R. Nalluri, "Evaluation of metrics and a dynamic thresholding strategy for high precision single sensor scene matching applications," *Multimedia Tools and Applications*, vol. 80, no. 12, pp. 18 803–18 820, 2021.

[16] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[17] H. Bay, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," in *European conference on computer vision*. Springer, 2006, pp. 404–417.

[18] J. Revaud, P. Weinzaepfel, C. De Souza, N. Pion, G. Csurka, Y. Cabon, and M. Humenberger, "R2d2: repeatable and reliable detector and descriptor," *arXiv preprint arXiv:1906.06195*, 2019.

[19] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proceedings of the 28th ACM international conference on Multimedia*, 2020, pp. 1395–1403.

[20] L. Ding, J. Zhou, L. Meng, and Z. Long, "A practical cross-view image matching method between uav and satellite for uav-based geo-localization," *Remote Sensing*, vol. 13, no. 1, p. 47, 2021.

[21] J. Zhuang, M. Dai, X. Chen, and E. Zheng, "A faster and more effective cross-view matching method of uav and satellite images for uav geolocalization," *Remote Sensing*, vol. 13, no. 19, p. 3979, 2021.

[22] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zhenga, and Y. Yang, "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[23] X. Tian, J. Shao, D. Ouyang, and H. T. Shen, "Uav-satellite view synthesis for cross-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[24] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A transformer-based feature segmentation and region alignment method for uav-view geo-localization," *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[25] H. O. Song, Y. Xiang, S. Jegelka, and S. Savarese, "Deep metric learning via lifted structured feature embedding," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4004–4012.

[26] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3961–3969.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

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

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[31] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *European conference on computer vision*. Springer, 2016, pp. 494–509.

[32] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays, "Learning deep representations for ground-to-aerial geolocalization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 5007–5015.

[33] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7132–7141.

[34] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha *et al.*, "Resnest: Split-attention networks," *arXiv preprint arXiv:2004.08955*, 2020.

[35] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[37] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 6105–6114.

[38] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-first AAAI conference on artificial intelligence*, 2017.

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[40] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2. IEEE, 2006, pp. 1735–1742.

[41] E. Hoffer and N. Ailon, "Deep metric learning using triplet network," in *International workshop on similarity-based pattern recognition*. Springer, 2015, pp. 84–92.

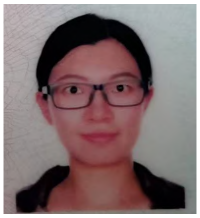
[42] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[43] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[44] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[45] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," *arXiv preprint arXiv:2111.09883*, 2021.

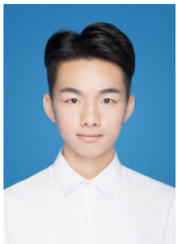
Runzhe Zhu received the B.S.degree in Zhejiang Shuren University from Zhejiang, Hangzhou, China, in 2020. He is currently a M.S.student with the department of electrical and electronic engineering of Shanghai University of Engineering Science, Shanghai, China. His research interests include vision localization and deep learning.



Ling YIN received the B.S. degree in software engineering from East China Normal University, China, in 2008, and the Ph.D. degree in Computer technology from East China Normal University, China, in 2016. She is currently a lecture with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, China. Her research interest includes deep learning, time series analysis, and software engineering with formal methods.

Mingze Yang received the B.S.degree in Rolling Stock Engineering from Shanghai Institute of Technology, Shanghai, China, in 2019. He is currently a M.S.student with the department of electrical and electronic engineering of Shanghai University of Engineering Science, Shanghai, China. His research interests include deep learning, target recognition and wireless sensing.

Fei Wu received the B.S. degree, the M.S. degree and the Ph.D. degree in Computer Science from National University of Defense Technology in 1990, 1993 and 1998. He was a PostDoctoral Research with Nankai University, China. He is currently a full professor with the School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, China. His research interests include intelligent information processing, positioning technology and machine learning.



Yuncheng Yang received the B.S.degree in mechanical engineering from Henan University of Science and Technology, Luoyang, China, in 2018. He is currently a M.S.student with the department of electrical and electronic engineering of Shanghai University of Engineering Science, Shanghai, China. His research interests include deep learning, indoor positioning and wireless sensing.

Wenbo Hu received the Ph.D. degree in Geography from Université Grenoble Alpes, Grenoble, France, in 2019. He is currently the post-doctor in School of Communication and Information Engineering, Shanghai University. From 2019 to 2020, he was a postdoctoral researcher with the Laboratoire PACTE, UMR 5194 CNRS, France. His research interest includes the behavioral geography, spatial modeling and public policing based on big data, deep learning and machine learning.