IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 5, 2022, accepted March 18, 2022, date of publication March 28, 2022, date of current version April 1, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3162693

A Semantic Guidance and Transformer-Based Matching Method for UAVs and Satellite Images for UAV Geo-Localization

JIEDONG ZHUANG[1], XURUOYAN CHEN[1], MING DAI[1], WENBO LAN[2], YONGHENG CAI[2], AND ENHUI ZHENG[1]

[1]Unmanned System Application Technology Research Institute, China Jiliang University, Hangzhou 310018, China
[2]China Academy of Aerospace Aerodynamics (CAAA), Beijing 100074, China

Corresponding author: Enhui Zheng (ehzheng@cjlu.edu.cn)

ABSTRACT It is a challenging task for unmanned aerial vehicles (UAVs) without a positioning system to locate targets by using images. Matching drone and satellite images is one of the key steps in this task. Due to the large angle and scale gap between drone and satellite views, it is very important to extract fine-grained features with strong characterization ability. Most of the published methods are based on the CNN structure, but a lot of information will be lost when using such methods. This is caused by the limitations of the convolution operation (e.g. limited receptive field and downsampling operation). To make up for this shortcoming, a transformer-based network is proposed to extract more contextual information. The network promotes feature alignment through semantic guidance module (SGM). SGM aligns the same semantic parts in the two images by classifying each pixel in the images based on the attention of pixels. In addition, this method can be easily combined with existing methods. The proposed method has been implemented with the newest UAV-based geo-localization dataset. Compared with the existing state-of-the-art (SOTA) method, the proposed method achieves almost 8% improvement in accuracy.

INDEX TERMS Cross-view image matching, geo-localization, UAV image localization, deep neural network.

I. INTRODUCTION

The researches on remote sensing images have been a hot topic for a long time. There is a part of research devoted to detecting targets from remote sensing images [1]–[4]. Some other works were dedicated to semantic segmentation of remote sensing images [5]–[8]. Another line of works focused on the large scene images classification [9]–[12]. In recent years, the booming development of Unmanned aerial vehicles (UAVs) has promoted the application of drones in all walks of life. UAV is easy to operate and shoot, which makes it popular and gradually become the main tool for acquiring remote-sensing images. So far, most drones on the market rely on positioning systems (e.g. GPS or GNSS) for positioning and navigation. Cross-view geo-localization is matching the drone images with the satellite images marked with geographic location, which makes drone to obtain the current position and realize autonomous positioning without the assistance of the positioning system.

Cross-view geo-localization is mainly to achieve image positioning by matching images from different perspectives, which is challenging because the appearance and view-points are significantly different in various views. Some previous work used some hand-designed features such as semantically labeled regions and feature translation for similarity calculations [13]–[16]. With the rapid development of deep learning and CNN, the method of manual features has been replaced by neural networks autonomously extracting features. One line of works focused on matching ground and satellite images, and implemented the method on the two datasets CVUSA [17] and CVACT [18]. In these two datasets, there is an image pair, containing a panoramic ground image and a satellite image for a location. Hu *et al.* [19] proposed a Siamese architecture to do metric learning for the matching task, which used NetVLAD [20] to encode local feature into global image descriptors. Liu and Li [18] designed a Siamese network which explicitly encodes the orientation

The associate editor coordinating the review of this manuscript and approving it for publication was Maurizio Magarini.

of each pixel in the images, significantly boosting the discriminative power of the learned deep features. Furthermore, Liu *et al.* [21] proposed a new Stochastic Attraction and Repulsion Embedding (SARE) loss function to minimize the gap between the learned and the actual probability distributions. Vo and Hays [22] proposed a new loss function which significantly improves the accuracy of Siamese and Triplet embedding networks and proved the effectiveness of orientation supervision. Shi *et al.* [23] applied polar transform to warp aerial images to align aerial and ground views. They also designed a DSM method [24] by adopting a dynamic similarity-matching network to estimate cross-view orientation alignment during localization. Another line of works focused on matching drone and satellite images, and implemented the method on the UAV-based datasets University-1652 [25]. Zheng *et al.* [25] looked at image-retrieval tasks from a classification perspective and optimized model by cross-entropy loss and instance loss [26]–[29]. Ding *et al.* [30] proposed a data augmentation method to solve the problem of unbalanced drone and satellite image samples in the dataset. In-spired by success of partition strategies [31]–[34] in other fields, Wang *et al.* [35] proposed a rotation-invariant square-ring feature partition strategy to enable the network to fully mining contextual information.

Since transformer was proposed by Vaswani *et al.* [36] in the field of NLP, it has maintained a high degree of popularity in deep learning. In recent years, excellent research based on transformer in the Computer Vision (CV) field has emerged one after another. With its unique self-attention mechanism and high performance, it may even replace CNN's long-standing dominance in CV. At present, transformer has penetrated into many subfields of CV, such as Image Classification, Object Detection, Semantic Segmentation and GAN, etc. [37]–[46].

In the field of Image Classification, Dosovitskiy *et al.* [37] proposed a pure transformer network, called Vision Transformer (ViT), which sequences image patches and performs very well on image classification tasks. Based on ViT, Touvron *et al.* [38] introduced a teacher-student strategy specific to transformers, called Data-efficient image Transformers (DeiT). DeiT used a distillation token to ensure the student learns from the teacher through attention, thereby speeding up the speed of network training and reducing the dependence on the amount of data. In the field of Object Detection, Carion *et al.* [39] removed the need for many hand-designed components and designed a transformer encoder-decoder architecture named Detection Transformer (DETR), which reasons about the relations of the objects and the global image context. However, DETR has some shortcomings such as slow convergence speed and limited feature spatial resolution. To solve these problems, Zhu *et al.* [40] proposed Deformable DETR, whose attention modules only at-tend to a small set of key sampling points around a reference. Inspired by transformer, Chi *et al.* [41] presented an attention-based decoder module to bridge various representations into a typical object detector. In the other fields of CV, there are also many excellent jobs. Zheng *et al.* [42] deployed a pure transformer called Segmentation Transformer (SETR) to encode an image as a sequence of patches and modeled global context in every layer of the transformer. Chen *et al.* [43] developed a new pre-trained model, named image processing transformer (IPT). IPT could be efficiently employed on low-level computer vision task (e.g. denoising, super-resolution and deraining). Jiang *et al.* [44] used two pure transformers to build a Generative Adversarial Network (GAN). To restore the texture information of the image super-resolution result, Yang *et al.* [45] proposed a novel Texture Transformer Network for Image Super-Resolution (TTSR) consisting of a learnable texture extractor by DNN, a relevance embedding module, a hard-attention module for texture transfer, and a soft-attention module for texture synthesis. He *et al.* [46] proposed a ViT-based pure transformer structure for pedestrian re-identification with embedding side information and jigsaw patch modules which improve discrimination ability of feature.

However, there is few transformer-based method that can be used for cross-view matching at present and the existing methods for extracting contextual information from drone and satellite images are only at the block level instead of the pixel level, which are not robust enough to offset and scale. To fill the above gaps, we mainly made the following contributions:

1. Different from other existing CNN-based methods, a Swin-transformer-based structure is proposed to match UAV and satellite images (see Sections II.B).

2. A semantic guidance module was proposed and used to realize the feature alignment of contextual information mining and inference stage improving the accuracy of the model under offset and scale (see Sections II.C).

3. The method achieved outstanding performance. On various accuracy indicators of the benchmark dataset, the method greatly exceeded the existing methods (see Section III).

The rest of this paper is organized as follows. In Section II, the methodology and materials are briefly introduced. Section III presents the experiments and results of our method. Discussion and conclusion are illustrated in Section IV and Section V respectively.

II. METHODOLOGY AND MATERIALS

A. DATASETS AND EVALUATION INDICATORS

The research was implemented with University-1652, released by Zheng *et al.* [25]. There are 1652 geographic targets in 72 universities from all over the world. The dataset of each target consists of images from three different perspectives, including satellite, drone, and street views. Each target has only one satellite-view image, about fifty drone-view images from different filming angles and heights, and some street-view images. This research focuses on the matching of satellite and drone views. The performance of method is mainly reflected in two tasks, Drone → Satellite and Satellite → Drone. Specifically, the purpose of Drone → Satellite is giving a drone image and finding the satellite

image of the same place; the purpose of Satellite → Drone is giving a satellite image and finding all drone images of the corresponding place. The details of data distribution in the datasets are shown in Table 1. In the testing dataset of the Drone → Satellite task, there was only one true-matched satellite-view image for each drone-view image.

All geotags satellite-view images were captured from Google Maps, which have a similar scale to that of drone-view images and high spatial resolutions (from level 18 to 20, the spatial resolution ranges from 1.07 to 0.27 m).

Due to airspace control and high cost, it is very difficult to collect a large number of real drone-view images, so the drone-view images were simulated by the 3D model provided by Google Earth. The view in the 3D model spirally descends, and the height of view from 256 to 121.5 m, while images were recorded at regular intervals, so as to obtain a large number of drone images close to the real world. As shown in Fig. 1, the blue curve represents the flight trajectory of the drone, and the blue cylinder represents the shooting target.

TABLE 1. Statistics experimental data.

Training Dataset		
Views	Numbers of buildings	Numbers of images
Drone	701	37,854
Satellite	701	701
Testing Dataset		
Views	Numbers of buildings	Numbers of images
$\text{Drone}_{\text{query}}$	701	37,854
$\text{Satellite}_{\text{query}}$	701	701
$\text{Drone}_{\text{gallery}}$	951	51,355
$\text{Satellite}_{\text{gallery}}$	951	951

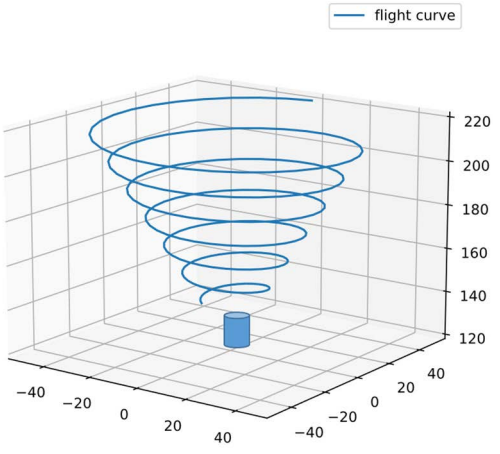

FIGURE 1. Data collection diagram.

To evaluate the performance of the proposed method, Recall@K (R@K) [47] and average precision (AP) [48] are selected as evaluation indicators which are two widely used measurements. Recall@K (R@K) represents the probability that a correct match appears in the top-k ranked retrieved results. And another metric, average precision (AP) measures the average retrieval performance with multiple ground truths, it is originally widely used in image retrieval.

B. OVERVIEW OF NETWORK

Difference from the other existing three branches method [25], [30], [35] is that the proposed network is composed of drone and satellite branches without street branch. In addition, the backbone is not traditional CNN structure (e.g. Resnet [49], VGG [50]), but is a new transformer structure Swin-Tiny [51], which has achieved good performance in many other computer vision fields. Swin-Tiny consists of 4 layers. Each layer contains 2, 2, 6 and 2 self-attention modules respectively. The structure of the self-attention module is shown in Fig. 2. In order to compare with other methods, schematically taking an image with the same size of 256×256 as that of the input of the network, the overview of network and the forward propagation process are shown in Fig. 2. The whole structure was divided into two branches, the drone and satellite view branches, and they share the weights of the backbone. The image is divided into 4 patches and sent to back-bone. A feature map with a size of 64×768 is obtained through Layer1 to Layer4. After being processed by Semantic Guidance Module (details in Section II.C), the feature map is split into several parts. Each part represents different semantics. Average pooling operation are then performed on each part, since the size of part become 1×768. All pooled features are sent to the classifier module, including fully connected, batch normalization, dropout, and classification layers. The network is optimized by minimizing cross-entropy loss in training phase. The classification layer in classifier module will be removed in inference (details in Section II.D).

C. SEMANTIC GUIDANCE MODULE (SGM)

Contextual information is critical to the accuracy of image retrieval. The existing hard partition strategy is not robust enough to offset and scale. Since a pixel-based partition strategy is proposed, which shows good performance in ablation experiments on offset and scale (details in Section III.C).

The input feature map of SGM can be expressed as M_i^j, and the size of M is 64×768. SGM sums M_i^j along the channel direction. The operation can be formulated as:

$$M_i = \sum_{j=0}^{768} M_i^j \quad i \in [0, 63] \tag{1}$$

After the above operations, the size of M_i is 64×1.

Normalization operation is performed on M_i, and the result is shown in Fig. 3a. The normalization operation can be formulated as:

$$M_i = \frac{M_i - Minimum(M_i)}{Maximum\,(M_i) - Minimum(M_i)} \tag{2}$$

where *Maximum* and *Minimum* respectively stand for getting the maximum and minimum in M_i.

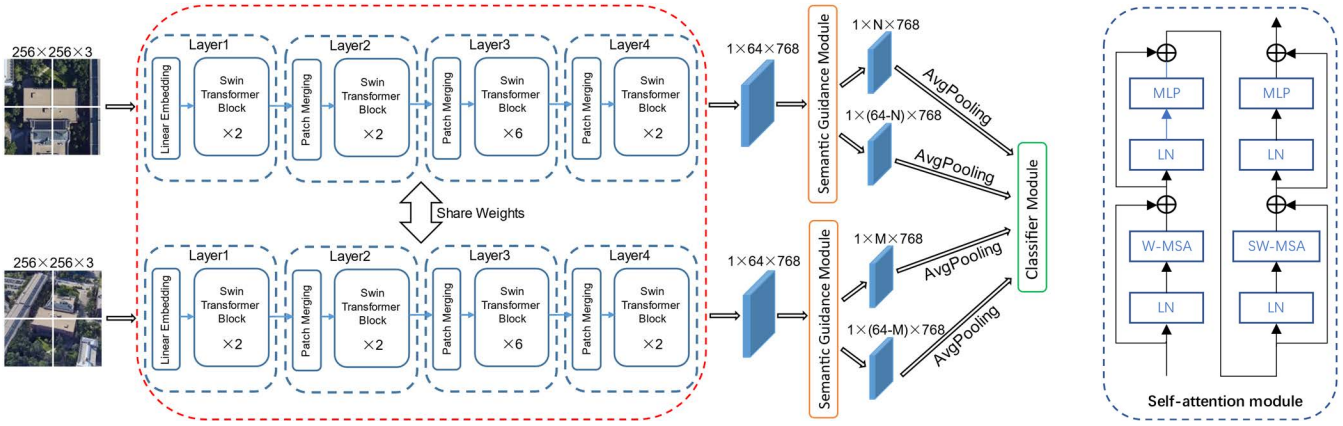

FIGURE 2. Overview of the network and the structure of the self-attention module. In the self-attention module: "LN" means layer normalization; "W-MSA" and "SW-MSA" mean multi-head self-attention modules with regular and shifted windowing configurations, respectively. "MLP" means multilayer perceptron.

Next SGM calculates the gradient between adjacent positions, and divides the feature map into different regions based on the calculation results. Take dividing the feature map into two parts as an example, the red arrow in Fig. 3a indicates a position with a large gradient. SGM regards this position as the line to divide the feature map into yellow and green parts (shown in Fig. 3b), which represent two categories in the image. The whole operation process can be formulated as:

$$i_{position} = argmax\left(\frac{M_{i+1} - M_i}{M_i}\right) \tag{3}$$

The overview and the result of SGM is shown in Fig. 4. After SGM, the feature map is obviously divided into two parts: architecture (foreground) and environment (background). This result is a good foundation for extracting contextual information. And we believe this partition method is robust to offset and scale.

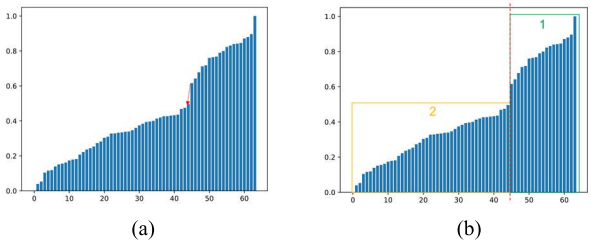

FIGURE 3. Values in the heatmap and the principle of SGM.

D. LEARNING STRATEGY AND LOSS FUNCTION

Average pooling is performed on each divided feature map, and their size is reduced to 1×768. The operation can be formulated as:

$$Y_i = Avgpool\ (X_i) \quad i \in [0, 1] \tag{4}$$

where X_i stands for divided feature map, $Avgpool$ stands average pooling operation and Y_i stands pooled vector.

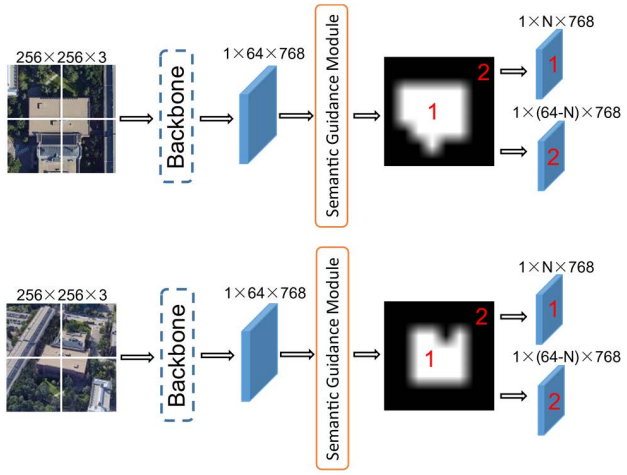

FIGURE 4. Overview of SGM.

All pooled vector is sent to classifier module, which includes fully connected, batch normalization, dropout, and classification layers. Each vector output by module is performed softmax to normalize the result to a feature space with the value range of 0 to 1. The optimization goal is that, in this feature space, the feature vectors of the same geographic target have a closer distance, on the contrary, the feature vectors of different geographic targets have a longer distance. The network is optimized by cross-entropy loss function and it can be formulated as:

$$Loss_{CE}(p, y) = \begin{cases} -\log\ (p), & y = 1 \\ -\log\ (1-p), & y = 0 \end{cases} \tag{5}$$

where p stands for forecast result, and y stands ground-truth label.

The total loss in the training phase can be formulated as:

$$Loss = \sum_{i=0}^{n} (L^i_{Drone} + L^i_{Satellite}) \tag{6}$$

where n stands for number of divided feature maps.

TABLE 2. Comparison with the state-of-the-art results reported on university-1652.

Method	Publication	Resolution	Backbone	Speed	Params	FLOPs	Drone→Satellite		Satellite→Drone	
							R@1	AP	R@1	AP
Soft Margin Triplet Loss [18]	CVPR'19	256×256	VGG16	1.39×	138M	16G	53.21	58.03	65.62	54.47
University-1652[25]	ACM MM'20	256×256	ResNet-50	-	26M	3.5G	58.49	63.13	71.18	58.74
Instance Loss [28]	TOMM'20	256×256	ResNet-50	-	26M	3.5G	58.23	62.91	74.47	59.45
Instance Loss + Verification Loss [29], [21]	TOMM'17	256×256	ResNet-50	-	26M	3.5G	61.30	65.68	75.04	62.87
Instance Loss + Gem Pooling [52]	TPAMI'18	256×256	ResNet-50	-	26M	3.5G	65.32	69.61	79.03	65.35
LCM [30]	Remote	256×256	ResNet-50	-	26M	3.5G	66.65	70.82	79.89	65.38
LPN [35]	TCSVT'21	256×256	ResNet-50	1.00×	26M	3.5G	74.16	77.39	85.16	73.68
LPN	TCSVT'21	224×224	ResNet-50	-	26M	2.7G	69.28	72.98	82.45	68.92
LPN	TCSVT'21	256×256	ResNet-101	1.51×	45M	7.3G	76.13	79.29	85.45	75.45
Ours	-	256×256	Swin-Tiny	1.04×	28M	5.9G	**82.14**	**84.72**	**88.16**	**81.81**
Ours	-	224×224	Swin-Tiny	-	28M	4.5G	79.59	82.50	87.73	79.59
Ours	-	224×224	Swin-Small	-	50M	8.7G	80.38	83.16	87.16	80.17
Ours	-	224×224	Swin-Large	-	197M	34.5G	85.44	87.60	90.16	85.28

In the testing phase, the classification layer in classifier module will be removed. Since the output size of module becomes 1×512. As shown in Fig. 5, two vectors are concatenated to calculate the Euclidean distance of the image in the feature space. The distance between two vectors can be formulated as:

$$D = \|V_{Drone} - V_{Satellite}\|_2 \tag{7}$$

where V stands for concatenated vector.

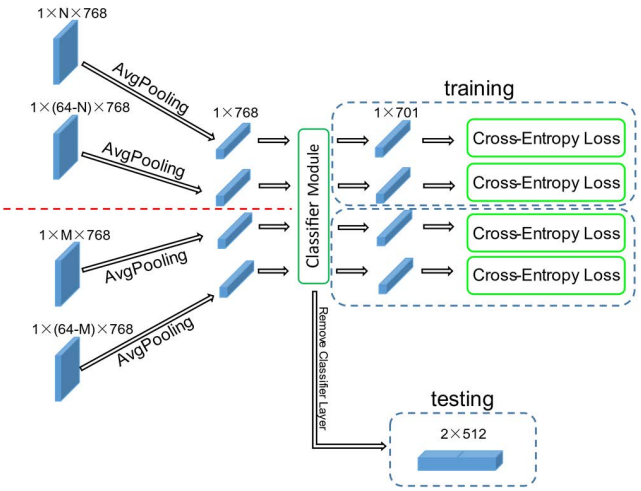

FIGURE 5. Training and testing stage.

E. IMPLEMENTATION

Training images were changed from 512×512 to 256×256 through the resize operation and augmented with random flipping and random cropping. The batchsize was set to 16. The backbone was initialized by loading the weights pretrained on ImageNet-1K by Swin-Tiny. SGD was chosen as the optimizer with momentum of 0.9 and weight decay of 5e-4 to train the model. The model was trained for 140 epochs; the initial learning rate was 9e−4 for the backbone layers, and 9e−3 for the other layers. The learning rate dropped to one-tenth of the original after 80 and 120 epochs. In the inference phase, the similarity between images was evaluated by calculating the Euclidean distance between L2-normalizing feature vectors of images. All experiments were performed with an Nvidia 3090 GPU using the PyTorch deep-learning framework with FP16 training.

III. EXPERIMENT

A. COMPARISON WITH THE STATE OF THE ART

In Table 2, the proposed method is compared with other methods on University-1652. The method has achieved 82.14 % R@1 accuracy and 84.72% AP on the task of Drone → Satellite, 88.16% R@1 accuracy and 81.81% AP on the task of Satellite → Drone with the standard input (image size of 256×256). The performance of the method greatly surpasses the existing competitive models. When choosing the model Swin-Large with a larger amount of parameters and calculations as backbone, which is pretrained on a larger datasets Imagenet-22K, the accuracy achieved higher. Experiments with different input sizes shows that input images with large resolution can obtain better matching accuracy in these two tasks.

B. ABLATION OF SGM

In order to verify the effectiveness of SGM, the distribution of embedding vectors after dimensionality reduction is visualized. As shown in Fig. 6, the SGM splits the vectors into

two parts as expected. To explore the impact of the number of parts divided by SGM, an ablation experiments are performed on SGM. As shown in Table 3 and Fig. 7, the method only achieved 79.19 % R@1 accuracy and 82.19% AP on the task of Drone → Satellite, 86.31% R@1 accuracy and 77.69% AP on the task of Satellite → Drone without SGM. When SGM was adopted and divided feature map into 2 parts, the method achieved 80.14% R@1 accuracy and 83.35% AP on the task of Drone → Satellite, 87.59% R@1 accuracy and 80.37% AP on the task of Satellite → Drone. When the number of parts increased to 3, model achieved the highest R@1 and AP. As the number of parts reached 4, the accuracy of the model dropped to 81.76 R@1 accuracy and 84.46% AP on the task of Drone → Satellite, 87.45% R@1 accuracy and 81.29% AP on the task of Satellite→ Drone. From the above phenomenon, we can judge that 3 is the best number of parts. Meanwhile, SGM brings 2.95%, 2.53%, 1.85% and 4.12% improvements on the 4 items in the table. Among them, the greatest improvement is the AP of the task of Satellite → Drone, which proves that SGM is an effective method. It makes satellite image retrieve more relevant drone images. Furthermore, in order to better integrate each channel of features, SE module [53] was inserted after each feature map output from SGM, which brought an objective improvement to the performance of the model.

TABLE 3. Results of ablation experiments of semantic guidance module. "2P-4P" means number of parts.

Method	Drone → Satellite		Satellite → Drone	
	R@1	AP	R@1	AP
baseline	79.19	82.19	86.31	77.69
baseline+SGM(2P)	80.44	83.35	87.59	80.37
baseline+SGM(3P)	82.14	84.72	88.16	81.81
baseline+SGM(4P)	81.76	84.46	87.45	81.29
baseline+SGM(3P)+SE	83.02	85.53	88.87	83.21

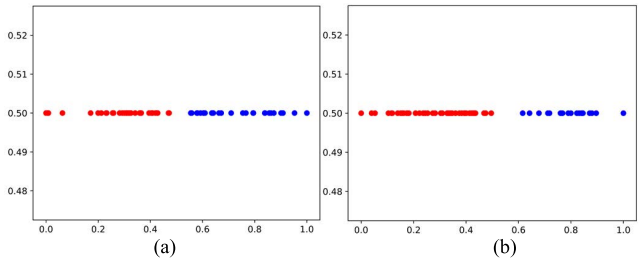

FIGURE 6. Examples of an embedding vector being divided into two parts by SGM.

C. ABLATION OF OFFSET AND SCALE

In order to confirm whether the model is robust to offset and scale or not a set of ablation experiments were designed. Firstly, the anti-offset of images was tested on the model. The query image was added by mirrored pixels at the edge of the image to achieve the effect of shifting the center target building. The operation process is shown in Fig. 8. The geographic target was shifted from 0 to 20 pixels from the center.

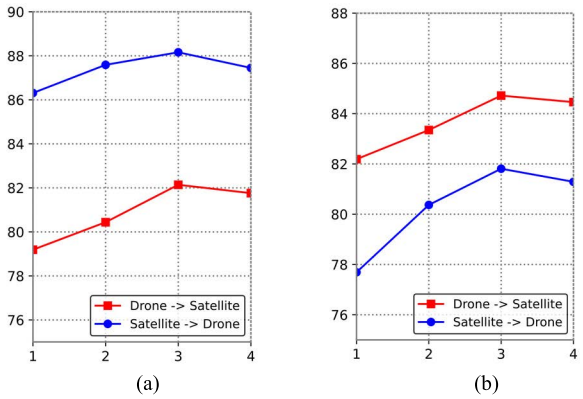

FIGURE 7. (a) The effect of the number of categories on R@1. (b) The effect of the number of categories on AP.

The experiment result is displayed in Table 4. The results showed that when the offset increased from 0 to 10, the model dropped to 81.20% R@1 accuracy and 84.07% AP on the task of Drone → Satellite, 87.87% R@1 accuracy and 80.51% AP on the task of Satellite → Drone. The model dropped less than 1% in various indicators. Even when the offset increased to 20, the model remained 79.26% R@1 accuracy and 82.22% AP on the task of Drone → Satellite, 85.88% R@1 accuracy and 79.23% AP on the task of Satellite → Drone, the decline of accuracy was less than 3%, which proves that the model was robust to the offset. Secondly, the robustness to scale of the model was tested on the task of Drone → Satellite. The drone-query images were split into three groups: short, medium and long, which respectively represented the different distances of the drone between the geographic target. As shown in Table 5, the model performs slightly worse on long distance, only achieved 79.92% R@1 accuracy and 83.05% AP. But on short and middle distance, the accuracy of the model exceeded the average level. We believe that it may be caused by closer scales between the middle distance images and satellite images. Nevertheless, there is no significant difference in the accuracy of the model on the three different distances, which indicates that the proposed model is robust to scale. The above two experiments show that the proposed method may adapt to complex situations in actual application scenarios.

D. INFERENCE WITH DIFFERENT PART

Each feature vector representing each part from SGM was use separately in inference phase to explore the effectiveness of contextual information extraction. Taking 3 parts in SGM as an example, the results of the ablation experiment are shown in Table 6. When each part was tested individually, they showed strong performance individually. Part3 (P3, the part with highest value in the feature map) even reached R@1

FIGURE 8. The way of shifting drone images.

TABLE 4. Results of ablation experiments of shifting query images during inference.

Padding Pixel	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
0	82.14	84.72	88.16	81.81
10	81.20	84.07	87.87	80.51
20	79.26	82.22	85.88	79.23

TABLE 5. Results of ablation experiments of using drone images with different distance to the geo-graphic target to conduct retrieval. "All" stands for using all drone-views query images.

Distance	Drone→Satellite	
	R@1	AP
All	82.14	84.72
Short	82.22	84.82
Middle	84.02	86.47
Long	79.92	83.05

80.35%, AP 83.31% on the Drone → Satellite task, and R@1 87.96%, AP 80.44% on the Satellite → Drone task. The results of the separation experiment prove that each part of the network had extracted effective context information. When all parts were combined, the model achieved the highest accuracy, R@1 82.14%, AP 84.72% on the Drone → Satellite task, and R@1 88.16%, AP 81.81% on the Satellite → Drone task. Compared with the individual inference of each part, the joint inference brought about 1% improvement in each indicator. The results of the joint inference experiment shows that the method of combining contextual information can improve the retrieval accuracy of the model. However, too much parts may cause an increase in inference time. P3 is a good choice in a scene with high real-time requirements.

E. MATCHING ACCURACY OF MULTIPLE QUERIES

In the above experiment of Drone → Satellite task, only a single drone view image was used to retrieve satellite view image, which was called "single mode". It is believed that a single drone view image can not provide complete information about geographic targets. To solve this problem, University-1652 provides a lot of drone images with different heights and angles for each geographic target, which provides convenience for us to retrieve satellite images based on the information of multiple drone images (called "multi mode"). In order to verify that multiple queries can improve retrieval accuracy, two sets of ablation experiments were designed.

TABLE 6. Results of ablation experiments of using drone images with different distance to the geo-graphic target to conduct retrieval.

Part Combination	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
P1	79.86	82.89	86.72	79.65
P2	79.85	82.85	87.01	79.31
P3	80.35	83.31	87.96	80.44
P1+P2+P3	82.14	84.72	88.16	81.81

TABLE 7. Results of multiple queries.

Method	Query mode	Drone→Satellite	
		R@1	AP
University-1652	Multi	69.33	73.14
LCM	Multi	77.89	81.05
Ours	Single	82.14	84.82
Ours	Multi	89.23	90.95

In the experiments the feature of multiple queries was set as the mean value of the single image features of a geographic target. Table 7 shows the accuracy of our proposed method in the two modes and comparison with other existing methods. When the "multi mode" was used on proposed method for retrieval, the model achieved 89.23% R@1 accuracy and 90.95% AP, which brought +7.09% R@1 and +6.13% AP to the model. Compared with other methods that use the "multi mode", the proposed method is also far ahead.

To further explore the impact of multiple angles and multiple heights on accuracy, an ablation experiment on angle and height were designed and the results were shown in Table 8. "Low", "Middle" and "High" are respectively represent drone images in different height ranges (shows in Fig. 9). When the height was kept in a certain range, using multi-angle drone images combination for retrieval will improve accuracy. The method that is called as "ALL" means to combine images of multiple heights on the basis of multiple angles. Compared with using a single height only, the accuracy of multiple heights had also been improved. Since we believe that multi-angle and multi-height queries both can improve the model accuracy.

F. EVALUATION ON REAL DATA

To verify that the method generalizes well, the model was evaluated on the dataset captured from our school. The model was trained on University-1652 and was not fine-tuned on this dataset. The dataset contains more than 50 locations and mainly focuses the accuracy of Drone → Satellite task. The evaluation results of different methods are shown in Table 9.

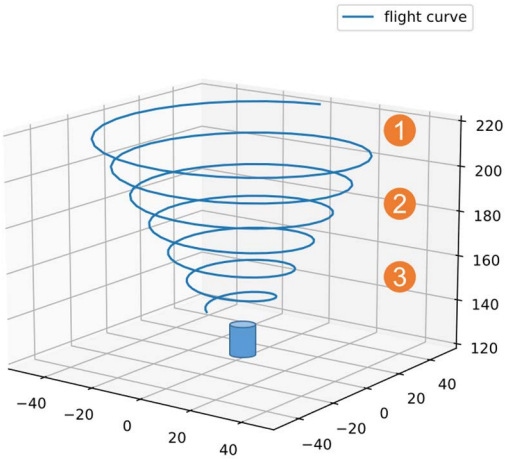

FIGURE 9. Flight curve of drone.

TABLE 8. Results of multiple queries about different flight height.

Flight height	Query mode	Drone→Satellite	
		R@1	AP
Low	Single	82.22	84.82
Middle	Single	84.02	86.47
High	Single	79.92	83.05
Low	Multi	85.67	88.20
Middle	Multi	88.38	90.39
High	Multi	87.95	89.88
All	Multi	89.23	90.95

TABLE 9. Results of inference on real datasets.

Method	R@1	AP
University-1652	7.51	7.57
LPN	35.26	31.03
Ours	54.65	53.49

G. VISUALIZATION OF RESULTS

The retrieval results of Satellite →Drone and Drone → Satellite tasks are displayed in Fig. 10 to prove the reliability of proposed method. Fig. 10a shows the method proposed had a high top-5 hit rate on Satellite → Drone, and Fig. 10b shows the model had a high R@1 accuracy.

To investigate the effectiveness of multiple queries, the matching results of multiple queries and a single query is visualized in Fig. 11. For the geographic target, the true-matched satellite image is in the position of R@5 when a single query is used. When multiple queries are used, the true-matched satellite image all appeared in the position of R@1. The result proves the effectiveness of the multiple queries. Fig. 12. shows the feature vectors of 11 classes images distributed in the Euclidean space. The same color represents the same class of images, it can be seen that the model has strong intra-class aggregation.

FIGURE 10. Visualized images retrieval results. (a) Top-5 retrieval results of Satellite → Drone. (b)Top-5 retrieval results of Drone→ Satellite. The true matches are in green boxes, while the false matches are displayed in red boxes.

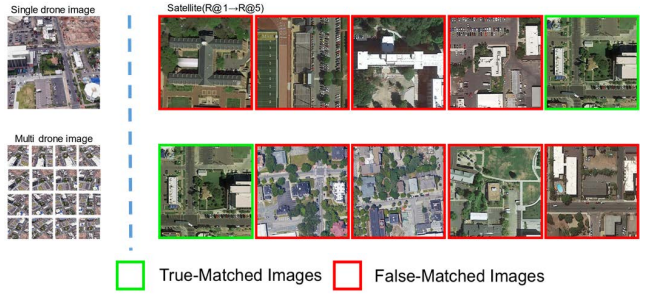

FIGURE 11. Visualized single query and multiple queries retrieval results.

IV. DISCUSSIONS

According to the experiments results on two tasks (Satellite → Drone and Drone → Satellite), we deeply explored the proposed model's retrieval performance and compared it with existing models. The baseline proposed using a transformer network as backbone showed a good performance without extra modules and tricks. As show in Fig. 13, the transformer can focus on more discriminative features in the

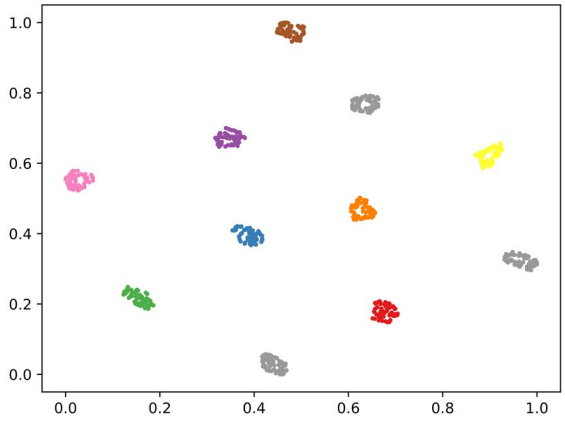

FIGURE 12. Visualization of partial data classification results using t-SNE.

FIGURE 13. Heatmaps generated by Vgg16, Resnet50 and Swin-T.

images than CNN. Because self-attention can explicitly mine the potential connections among patches in the whole images, but the convolution operation of CNN is more inclined to mine local features in the image. Since, compared with other complex networks with CNN as the backbone, transformer appears to be more competitive in image retrieval, which is conducive to the transformer being able to extract more characteristic features. Based on the strong baseline, SGM is proposed to extract richer contextual information and achieve feature alignment. When the feature map was divided into too many parts by SGM, the performance of the model declined. We conjecture that there are two reasons for this phenomenon: (1) Too many parts may cause redundancy of context information. (2) Too many branches will cause network overfitting. Therefore, it is necessary to choose an appropriate number of parts (e.g. 3 parts). In inference stage, if there are resource constraints, it is a good choice a branch with higher accuracy in SGM to reduce time expenditure. In order to achieve higher precision in real scenes, a method of multiple queries was used in the experiment. The results prove that multiple queries can significantly improve the accuracy of retrieval, which also guides us to allow drones to take multi-angle and multi-height shooting of the same geographic location to obtain more diverse features in the real scene. Nevertheless, it is believed that the model still has certain flaws. For example, the accuracy of SGM's semantic guidance in some complex scenes needs to be improved, which is also directly related to the final matching accuracy of the model. How to make better semantic guidance is meaningful and promising work, and we will conduct further research on this problem.

V. CONCLUSION

In the paper, we proposed a transformer-based network to match drones with satellite images, which can be used for drone autonomous positioning without a positioning system. A semantic guidance method is proposed to extract the contextual information in the image and improve the model's robustness to offset and scale. In the two tasks (Satellite → Drone and Drone → Satellite) on the dataset University-1652, the model achieved high accuracy.

The conclusion of the experiment are mainly as follows: 1. The transformer-based network is more competitive than the CNN-base network in this task. 2. The Semantic guidance module (SGM) can effectively mine the contextual information in the image, and achieve feature alignment in the inference stage, further improving the accuracy of the model. 3. Multiple queries is more accurate than single query, and it brings a huge improvement. The proposed method therefore achieved a precision that greatly surpasses existing methods.

ACKNOWLEDGMENT

The authors would like to thank Zheng Zhedong from the Reler Laboratory, University of Technology Sydney, for providing the Universty-1652 dataset, open source code, and his contribution in this field.

REFERENCES

[1] K. Fu, Z. Chang, Y. Zhang, G. Xu, K. Zhang, and X. Sun, "Rotation-aware and multi-scale convolutional neural network for object detection in remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 294–308, Mar. 2020.

[2] K. Li, G. Wan, G. Cheng, L. Meng, and J. Han, "Object detection in optical remote sensing images: A survey and a new benchmark," *ISPRS J. Photogramm. Remote Sens.*, vol. 159, pp. 296–307, Jan. 2020.

[3] Y. Zhao, L. Zhao, B. Xiong, and G. Kuang, "Attention receptive pyramid network for ship detection in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 2738–2756, 2020, doi: 10.1109/JSTARS.2020.2997081.
[4] M. Sharma, M. Dhanaraj, S. Karnam, D. G. Chachlakis, R. Ptucha, P. P. Markopoulos, and E. Saber, "YOLOrs: Object detection in multimodal remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1497–1508, 2021, doi: 10.1109/JSTARS.2020.3041316.
[5] Q. Wu, F. Luo, P. Wu, B. Wang, H. Yang, and Y. Wu, "Automatic road extraction from high-resolution remote sensing images using a method based on densely connected spatial feature-enhanced pyramid," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3–17, 2021, doi: 10.1109/JSTARS.2020.3042816.
[6] T. Mao, H. Tang, and W. Huang, "Unsupervised classification of multispectral images embedded with a segmentation of panchromatic images using localized clusters," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 11, pp. 8732–8744, Nov. 2019.
[7] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, 2020, doi: 10.1109/JSTARS.2020.3037893.
[8] X. Sun, A. Shi, H. Huang, and H. Mayer, "BAS^{4}Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5398–5413, 2020, doi: 10.1109/JSTARS.2020.3021098.
[9] M. Sheykhmousa, M. Mahdianpari, H. Ghanbari, F. Mohammadimanesh, P. Ghamisi, and S. Homayouni, "Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 6308–6325, 2020, doi: 10.1109/JSTARS.2020.3026724.
[10] G. Cheng, X. Xie, J. Han, L. Guo, and G.-S. Xia, "Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, no. 99, pp. 3735–3756, Jun. 2020, doi: 10.1109/JSTARS.2020.3005403.
[11] Y. Yuan and L. Lin, "Self-supervised pretraining of transformers for satellite image time series classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 474–487, 2021, doi: 10.1109/JSTARS.2020.3036602.
[12] Q. Sang, Y. Zhuang, S. Dong, G. Wang, H. Chen, and L. Li, "Improved land cover classification of VHR optical remote sensing imagery based upon detail injection procedure," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 18–31, 2021, doi: 10.1109/JSTARS.2020.3032423.
[13] F. Castaldo, A. Zamir, R. Angst, F. Palmieri, and S. Savarese, "Semantic cross-view matching," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop (ICCVW)*, Dec. 2015, pp. 1044–1052.
[14] T.-Y. Lin, S. Belongie, and J. Hays, "Cross-view image geolocalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 891–898.
[15] T. Senlet and A. Elgammal, "A framework for global vehicle localization using stereo images and satellite and road maps," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCV Workshops)*, Nov. 2011, pp. 2034–2041.
[16] M. Bansal, H. S. Sawhney, H. Cheng, and K. Daniilidis, "Geo-localization of street views with aerial image databases," in *Proc. 19th ACM Int. Conf. Multimedia (MM)*, 2011, pp. 1125–1128.
[17] M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs, "Predicting ground-level scene layout from aerial imagery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 867–875.
[18] L. Liu and H. Li, "Lending orientation to neural networks for cross-view geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5624–5633.
[19] S. Hu, M. Feng, R. M. H. Nguyen, and G. H. Lee, "CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7258–7267.
[20] R. Arandjelović, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "NetVLAD: CNN architecture for weakly supervised place recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2016, pp. 5297–5307.
[21] L. Liu, H. Li, and Y. Dai, "Stochastic attraction-repulsion embedding for large scale image localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2570–2579.
[22] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 494–509.
[23] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for image based cross-view geo-localization," in *Proc. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2019, pp. 8–14.
[24] Y. Shi, X. Yu, D. Campbell, and H. Li, "Where am I looking at? Joint location and orientation estimation by cross-view matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4063–4071.
[25] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 1395–1403.
[26] Y. Ho and S. Wookey, "The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling," *IEEE Access*, vol. 8, pp. 4806–4813, 2020, doi: 10.1109/ACCESS.2019.2962617.
[27] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022, doi: 10.1109/LGRS.2021.3098774.
[28] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 16, no. 2, pp. 1–23, May 2020.
[29] Z. Zheng, L. Zheng, and Y. Yang, "A discriminatively learned CNN embedding for person reidentification," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 14, no. 1, pp. 1–20, 2017.
[30] 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 Sens.*, vol. 13, no. 1, p. 47, Dec. 2020.
[31] Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang, "Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline)," in *Proc. Eur. Conf. Comput. Vis.*, Munich, Germany, Sep. 2018, pp. 501–518.
[32] L. Zheng, Y. Huang, H. Lu, and Y. Yang, "Pose-invariant embedding for deep person re-identification," *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4500–4509, Sep. 2019.
[33] L. Wei, S. Zhang, H. Yao, W. Gao, and Q. Tian, "GLAD: Global–local-alignment descriptor for scalable person re-identification," *IEEE Trans. Multimedia*, vol. 21, no. 4, pp. 986–999, Apr. 2019.
[34] Z. Zheng, L. Zheng, and Y. Yang, "Pedestrian alignment network for large-scale person re-identification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 10, pp. 3037–3045, Oct. 2018.
[35] T. Wang, Z. Zheng, C. Yan, J. Zhang, Y. Sun, B. Zheng, and Y. Yang, "Each part matters: Local patterns facilitate cross-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 867–879, Feb. 2022, doi: 10.1109/TCSVT.2021.3061265.
[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaise, and I. Polosukhin, "Attention is all you need," in *Proc. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 6000–6010.
[37] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is 11 worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, May 2021.
[38] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*.
[39] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2020, pp. 213–229.
[40] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *Proc. Int. Conf. Learn. Represent.*, May 2021.
[41] C. Chi, F. Wei, and H. Hu, "RelationNet++: Bridging visual representations for object detection via transformer decoder," in *Proc. Neural Inf. Process. Syst.*, Dec. 2020, pp. 6–11.
[42] S. Zheng, J. Lu, H. Zhao, X. Zhu, Z. Luo, Y. Wang, Y. Fu, J. Feng, T. Xiang, P. H. S. Torr, and L. Zhang, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," 2021, *arXiv:2012.15840*.
[43] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12299–12310.

[44] Y. Jiang, S. Chang, and Z. Wang, "TransGAN: Two pure transformers can make one strong GAN, and that can scale up," 2021, *arXiv:2102.07074*.
[45] F. Yang, H. Yang, J. Fu, H. Lu, and B. Guo, "Learning texture transformer network for image super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5790–5799.
[46] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11–17.
[47] X. Wang, G. Doretto, T. Sebastian, J. Rittscher, and P. Tu, "Shape and appearance context modeling," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.
[48] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.
[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015, pp. 7–9.
[51] 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 *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 11–17.
[52] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with, no., human annotation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1655–1668, Jun. 2018.
[53] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020, doi: 10.1109/TPAMI.2019.2913372.

JIEDONG ZHUANG received the B.S. degree in automation from China Jiliang University, Hangzhou, China, in 2019, where he is currently pursuing the M.S. degree with the Department of Control Engineering. His research interests include deep learning and image retrieval.

XURUOYAN CHEN received the B.S. degree in automation from China Jiliang University, Hangzhou, China, in 2019, where she is currently pursuing the M.S. degree with the Department of Control Engineering. Her research interests include image processing and simultaneous localization and mapping.

MING DAI received the B.S. degree in automation from China Jiliang University, Hangzhou, China, in 2020, where he is currently pursuing the M.S. degree in mechanical engineering. His research interests include deep learning and image processing.

WENBO LAN received the M.S. degree in control science and engineering from the Harbin Institute of Technology, China, in 2009. He is currently a Senior Engineer with China Academy of Aerospace Aerodynamics (CAAA). His research interests include UAV application and general design of UAV

YONGHENG CAI received the M.S. degree in control science and engineering from the Harbin Institute of Technology, China, in 2012. He is currently a Senior Engineer with China Academy of Aerospace Aerodynamics (CAAA). His research interests include UAV application, navigation flight control system design, and simulation of UAV.

ENHUI ZHENG received the Ph.D. degree in control science and engineering from Zhejiang University, Hangzhou, China, in 2006. He is currently an Associate Professor with the Department of Automation, China Jiliang University, Hangzhou. He is the Deputy Secretary-General of the Zhejiang Model Radio Sports Association. His research interests include UAV application, image processing, deep learning, and simultaneous localization and mapping.