IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 19 August 2023, accepted 16 September 2023, date of publication 22 September 2023,
date of current version 27 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317950

RESEARCH ARTICLE

Cross-View Geo-Localization for Autonomous UAV Using Locally-Aware Transformer-Based Network

DUC VIET BUI, MASAO KUBO, AND HIROSHI SATO
Department of Computer Science, National Defense Academy, Yokosuka, Kanagawa 239-0811, Japan
Corresponding author: Duc Viet Bui (vietviet2411@gmail.com)

This work was supported by the NEC C&C Foundation Grants for Researchers.

ABSTRACT Although GPS is commonly used for the autonomous flying of unmanned aerial vehicles (UAVs), researchers mainly focus on image-based localization methods due to their tremendous advantages when it comes to GPS-denied environments. In this study, we study the problem of image-based geo-localization between UAV and satellite (known as cross-view geo-localization), which is an essential step towards image-based localization. In cross-view geo-localization, extracting fine-grained features containing contextual information from images is challenging due to the large gap in visual representations between different views. Existing methods in this field often use convolutional neural networks (CNNs) as feature extractors. However, CNNs have some limitations in receptive fields, which leads to the loss of fine-grained information. Some researchers have implemented Transformer-based networks to overcome these circumstances. However, these approaches only focused on understanding the meaning of each pixel based on their attention and only partially utilized tokens that are produced from Transformer blocks. Therefore, different from these works, we proposed a Vision Transformer-based network that takes advantage of local tokens, especially the classification token. Through experiments, our proposed model has significantly outperformed existing state-of-the-art models, which gave promising capabilities for developing this method in the future.

INDEX TERMS Cross-view geo-localization, image retrieval, UAV localization, vision transformer, deep neural network.

I. INTRODUCTION

The problem of developing a completely autonomous UAV system has been a hot research topic in recent years due to its vast applications in various fields. For example, an autonomous UAV can fly into zones where humans can not pass or use radio-controlled vehicles to explore. The autonomous UAV would be convenient for rescue missions in dangerous places [1]. However, precise navigation technology is necessary to achieve a completely autonomous UAV system. Although Global Positioning System (GPS) is commonly used for localization and navigation of most autonomous UAVs, GPS tends to be inactive when operating in specific environments, such as urban canyons and dense foliage. Additionally, GPS can experience signal disruptions or outages due to jamming, interference, or adverse weather conditions, leaving UAVs without reliable positioning information [2].

Recently, image-based localization for autonomous UAVs has become famous because of the rich information from captured images and the rapid development of image processing, especially when deep learning has significantly improved the accuracy and reliability of these tasks. This field comprises two main approaches: relative visual localization (RVL) and absolute visual localization (AVL) [3]. RVL is often considered frame-to-frame localization, which focuses on understanding the camera's position and orientation relative to a previously observed or mapped environment. The RLV

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

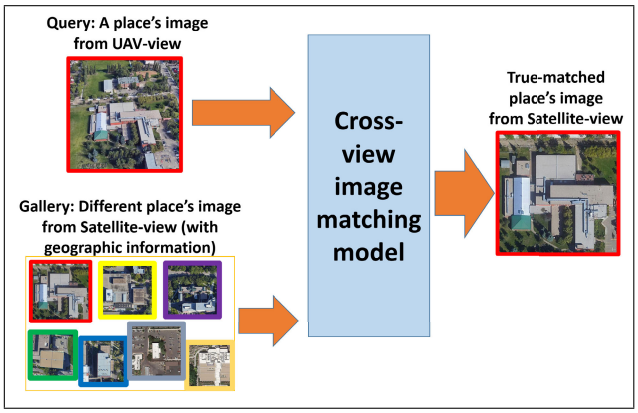


FIGURE 1. Example of cross-view image geo-localization (UAV - Satellite).

includes popular methods such as Visual Odometry (VO) [4] and Simultaneous Localization and Mapping (SLAM) [5], and many of them have been put into practice [6], [7] [8]. However, RVL is heavily affected by error accumulation: the error made in the prior pose estimation will impact the accuracy of the following estimation. Thus, the UAV tends to drift over time. This issue limits the capability of RVL in long-time applications and complex environments.

On the other hand, absolute visual localization (AVL), also known as frame-to-reference localization or geo-localization, [9], [10], involves estimating the geographical location of an image based on its visual content. This image-based geo-localization aims to determine the precise location from which an image was taken, typically using a previously collected database of geo-tagged images with known coordinates. In the context of autonomous UAVs, this research approach has been addressed as cross-view image matching for geo-localization (in short, cross-view geo-localization) [11], which is the task of matching the images from a UAV viewpoint with the satellite-view images annotated with geographic locations (UAV - Satellite). Figure 1 describes an example of this task. By finding the satellite-view image that is truly matched the captured image, the UAV can obtain the current geographic location and operate autonomous localization without GPS. In this paper, we mainly focus on alternative approaches to improve the accuracy and reliability of the task of cross-view geo-localization.

Early-stage of cross-view geo-localization-related works applied traditional image processing methods such as SURF [12] or SIFT [13] to create hand-crafted features, and after that, discriminate images by calculating the similarity score between these features. However, as the gap between different views is enormous, these methods could have gained better results than expected. In recent years, with the advances of deep learning in computer vision, hand-crafted feature-extracting methods have been replaced by learnable, autonomous feature-extracting methods such as convolutional neural networks (CNN). Thus, researchers in fields of cross-view geo-localization have achieved some remarkable results [14], [15] [16], [17]. However, the crucial key to solving cross-view geo-localization is finding relevant information between images and fully understanding global contextual information, which is quite challenging for CNNs as most CNN architectures often focus on small discriminative features.

In order to overcome these drawbacks of CNN-based methods in solving cross-view geo-localization, some works have brought to attention modules [18], [19] [20], [21] - which are famous for emphasizing the necessary parts and suppressing irrelevant parts in feature maps, thus, enhance the final contextual information. Additionally, with the evolution of self-attention mechanism in nature language processing, self-attention based network for vision processing - the Vision Transformer structure has been implemented in some cross-view geo-localization related works [22], [23], [24], [25] and achieved remarkable results. These methods used Vision Transformer as a robust contextual information extractor and processes input image at the pixel level (SGM [24]) or focused on the attention level of the local tokens which are embedded patches of input images (FSRA [25]). However, these methods still need to fully utilize classification tokens, which is also an essential component of the Vision Transformer. As the deeper and more layers the Transformer, the classification token can accumulate information from the other tokens in the sequence. It can effectively contribute to the performance of the Vision Transformer. Therefore, utilizing this token may unlock numerous potential in various applications. To investigate the effect of classification tokens in combination with other local tokens on the cross-view geo-localization problem, we mainly made the following contributions:

- We introduced a new Vision-Transformer-based architecture, which used a token enhancement strategy combining classification tokens and local tokens to improve the accuracy of matching UAV and satellite images.
- Our proposed model has shown outstanding performance through several extensive experiments on the benchmark dataset (University-1652) and greatly exceeded the current state-of-the-art methods.

The rest of the paper is organized as follows. Section II outlines the related research conducted in this field. In Section III, we introduce our proposed model. Section IV demonstrates the results of the experiments and some discussions. Finally, Section V presents a summary and our future challenges.

II. RELATED WORKS

In this Section, we review some famous approaches in the field of cross-view geo-localization (Section II-A) and introduce the applications of Transformer in computer vision (Section II-B), along with recent token enhancement strategies of Vision Transformer (Section II-C).

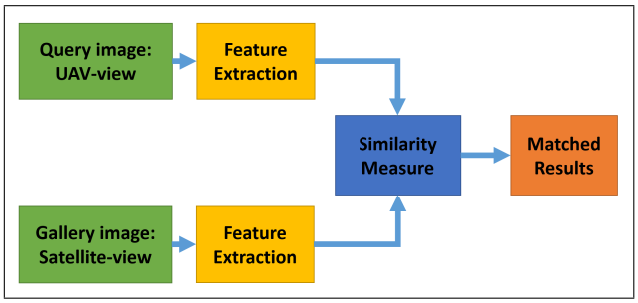


FIGURE 2. The pipeline of cross-view geo-localization.

A. CROSS-VIEW IMAGE MATCHING FOR GEO-LOCALIZATION

Previous works on cross-view geo-localization often focus on two tasks: the task of matching panoramic street-view images and satellite images (which were usually conducted on CVUSA [26] and CVACT [27] datasets) and the task of matching UAV-view images and satellite-view images (which were conducted on University-1652 [11] dataset). These works considered cross-view image matching for geo-localization as an image-retrieval task: giving an image, the matching model needs to retrieve the true-matched image from a different view's gallery. The pipeline of cross-view geo-localization (UAV $\rightarrow$ Satellite) can be described in Figure 2. At first, features from both query and gallery images are extracted using different feature processing methods. After that, the similarity of both query and gallery features is computed by standard similarity measures such as Cosine similarity or Euclidean distance [28]. These results are added to a ranking list, later used to find the true-matched image pair. Finally, the geo-tag in the true-matched gallery image is used for the next phase of localization.

In some existing works, hand-crafted features extracted by traditional feature processing methods have been applied [9], [29], [30]. The development of cross-view geo-localization continued to grow when deep learning methods were introduced in this field. Feature learning-based models (CNNs) and Metric learning methods are typical for learning image representations. For example, in matching street-view and satellite images, Workman et al. [31] adopted deep networks for the first time in cross-view image matching and achieved remarkable results. Hu et al. [14] designed a model that applied NetVLAD - a famous CNN model in image retrieval, with Siamese network architecture and trained it with metric loss. Moreover, Liu et al. [27] proposed a Siamese network to encode the orientation information of each pixel in the images. Vo and Hays [32] proposed soft-margin ranking loss in cross-view matching to overcome the margin issues of margin triplet loss, and Hu et al. [14] further improved this loss by introducing weighted soft-margin ranking loss, which significantly reduced convergence in the training phase. In the task of matching UAV-view images and satellite-view images, Zheng et al. [11] proposed ResNet-based models and optimized them with cross-entropy loss and instance loss [33]. Ding et al. [34] considered this problem a place classification task and solved it with an image classification model. LPN [35] and MBSA [36] derived the idea of utilizing part-based features from PCB [37] and proposed a square-ring feature partition strategy, which enabled networks to exploit small contextual information fully.

B. TRANSFORMER IN COMPUTER VISION

The attention mechanism was first proposed as an effective technique to help neural networks find out the most critical region of an image, thus increasing the model performance in feature learning. For several years, attention mechanism has been deployed in various computer vision tasks, including cross-view geo-localization. Recently, the power of the self-attention mechanism in Natural Language Processing, Transformer, has gained lots of attention from researchers. Some of them tried to utilize its performance in vision processing-related tasks. In 2020, Dosovitskiy et al. [38] was the first to apply a Transformer in the image classification task, known as Vision Transformer (ViT). Since then, Transformer has been widely developed in many mainstream vision processing problems, such as object detection, semantic segmentation, and Person Re-Identification (Person Re-ID), and has received a lot of remarkable results. For example, the Vision Transformer has shown promise in object detection tasks, which aim to locate and classify objects within an image. Detection Transformer (DETR) [39] was the first Transformer-based network for object detection tasks; Zhu et al. [40] proposed the DeepViT, a model which incorporates Vision Transformers into object detection frameworks like YOLO and Faster R-CNN, showcasing superior performance on object detection benchmarks.

Additionally, the self-attention mechanism in the ViT also allows the model to effectively capture global context, leading to more precise and contextually aware segmentation results. Segmentation Transformer (SETR) [41] was the first pure-transformer model for semantic segmentation; Carion et al. [42] extended the ViT's capabilities to semantic segmentation, where they utilized a hybrid architecture combining ViT with the DETR object detection model, showcasing significant improvements over previous methods. In the field of Person Re-ID, TransReID [43] successfully utilized the power of Transformer to achieve competitive results compared to CNN-based methods; Lu et al. [44] designed an end-to-end dual-branch Transformer network for occluded person re-identification, which surpassed the state-of-the-art results on benchmark Person Re-ID datasets.

For the cross-view geo-localization task, some researchers started implementing Transformer-based methods for the ground-satellite geo-localization. Yang et al. [22] proposed a self-cross attention Transformer network to learn the representations of both ground and satellite views. Zhu et al. [23] designed TransGeo that utilizes the strengths of Transformer related to global information modeling

and explicit position information encoding, which achieves state-of-the-art results on several datasets and takes less computation cost than CNN-based methods. Tian et al. [45] proposed a cross-view matching method called SMDT with a new image alignment strategy combined with Transformer, which is superior to existing methods. However, there are only a few applications of Transformer in UAV-satellite image matching problems. Inspired by TransReID, Dai et al. [25] designed FSRA. This network automatically divides regions based on the heat distribution of the Transformer's feature maps and aligns them in different views one on one. Zhuang et al. [24] also shared the same idea with FSRA but used a Swin-Transformer [46] backbone aligned the same semantic parts of two images by classifying each pixel based on the attention value of pixels.

C. LOCAL TOKEN ENHANCEMENT

The architecture of Vision Transformer by Dosovitskiy et al. [38] composed of several Multi-Head Self-Attention (MHSA) blocks that operate self-attention mechanism on embedded patches of input images (known as tokens), which are later used in Multi-Layer Perceptron (MLP) to perform final classification. After the debut of Vision Transformer, most Transformer-related studies focused on enhancing the MHSA block [46], [47], [48], while some researchers paid attention to the correspondence between these tokens. Beal et al. [49] was the first to combine local tokens to create spatial feature maps for object detection. After that, Jiang et al. [50] proposed token labeling - a new training strategy that labels tokens based on their attribution to take advantage of all patch tokens. Yuan et al. [51] designed Tokens-to-Tokens Vision Transformer (T2T-ViT), which introduced the tokens-to-token (T2T) process to tokenize images to tokens progressively and structurally aggregate tokens.

In this work, our approach takes inspiration from FSRA [25], which combines local tokens using a jigsaw classification branch, and especially the Locally-Aware Transformer (LA-Transformer) from the work of Sharma et al. [52]. LA-Transformer combined classification token with the local tokens and re-arranged them in the form of a 2D image grid like the strategy from PCB [37] to take advantage of the 2D spatial locality of these tokens, which performed successfully in the Person Re-ID task. Both person Re-ID task and cross-view geo-localization task used the same approach of representation learning - which encouraged us to apply this idea of LA-Transformer on a cross-view geo-localization task.

III. PROPOSED METHOD

In Section III, we explain the details of our proposed network: overview of the proposed network architecture (Section III-A); the structure of the Vision Transformer backbone (Section III-B); the Token Enhancement Process (Section III-C); the Classifier Module (Section III-D) and Loss Function (Section III-E).

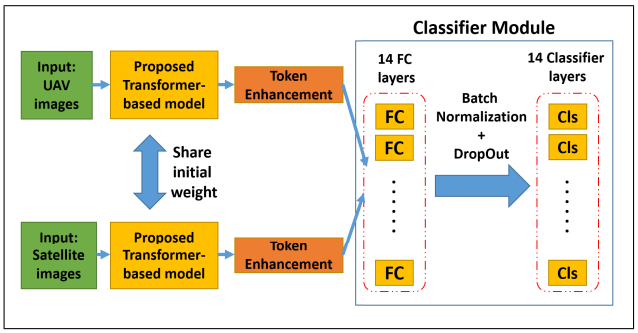


FIGURE 3. Proposed network architecture.

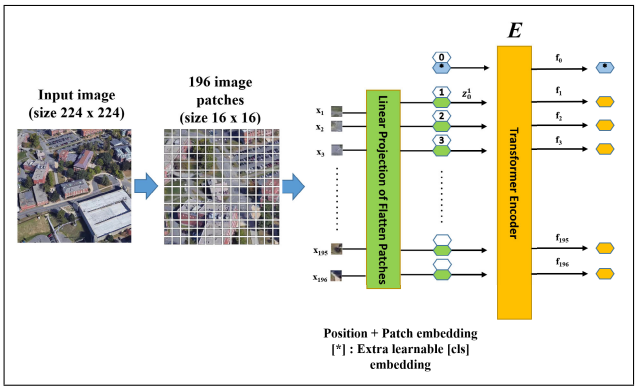


FIGURE 4. Vision transformer backbone.

A. OVERVIEW OF PROPOSED NETWORK ARCHITECTURE

Figure 3 demonstrated our proposed network architecture overview. The network consists of two branches. One branch takes UAV-view images as input, and the other takes satellite-view images as input. For the backbones, we apply a basic Vision Transformer (ViT) on both branches, and they all share initial weights that were pre-trained on the ImageNet dataset. In both branches, we employ a token enhancement strategy. All the feature outputs are transferred to a classifier module composed of Fully-Connected (FC) layers and Classifier layers (Cls), with Batch Normalization and DropOut.

B. VISION TRANSFORMER BACKBONE

Figure 4 described the structure of the ViT backbone. Given an input $x \in \mathbb{R}^{H\times W\times C}$, the model divided the input image into N number of small flattened 2D patches:

$$x_p^i | i = 1, 2, \cdots, N \tag{1}$$

where $x_p^i \in \mathbb{R}^{K^2\cdot C}$, (H, W) is the resolution of the original image, C is the number of channels, (K, K) is the resolution of each image patch. The number of patches N is calculated as follows:

$$N = \frac{HW}{K^2}. \tag{2}$$

Here we used the original ViT-B from [38], which has an input size of 224×224 and an original patch size of 16×16. Thus, the number of patches N will be 196. After that, the Linear

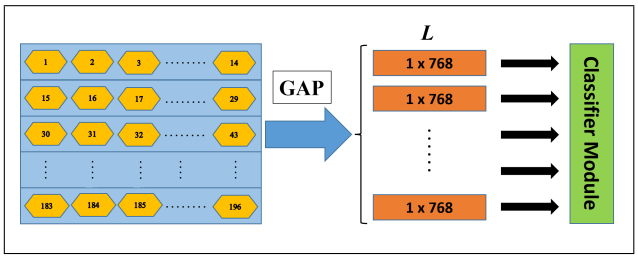


FIGURE 5. The token enhancement strategy.

Projection Flatten Patches linearly converted these patches into D dimensions using patch embedding function E:

$$E(x_p^i)|i = 1, 2, \cdots, N. \tag{3}$$

Here, D is set to 768. Before transferring these patches to the Transformer Encoder, an extra learnable embedding token x_{cls}, whose state at the output of the Transformer encoder serves as the image representation, is added, and all the tokens are fused with position embedding P to preserve the positional encoding information. The final vector z_0, which composed of N number of patches and class embedding x_{cls} can be defined as follows:

$$z_0 = [x_{cls}; E(x_p^1); E(x_p^2); \ldots; E(x_p^N)] + P \tag{4}$$

After that, the final vector z_0 was transferred through Transformer Encoder F, which consists of multiple Transformer Blocks. The final output contained $N + 1$ feature vectors (in this study, these are called tokens) and can be defined as follows:

$$F(z_0) = [f_0, f_1, f_2, \cdots, f_N] \tag{5}$$

C. TOKEN ENHANCEMENT PROCESS

In this study, the token enhancement process follows the work of LA-Transformer by Sharma et al. [52], which was also inspired by the PCB technique of Sun et al. [37]. As explained above, the Transformer encoder outputs $N + 1$ tokens:

$$F(z_0) = [f_0, f_1, f_2, \cdots, f_N]. \tag{6}$$

Here, we denote global token (classification token) as $G = f_0$ and local tokens as $Q = [f_1, f_2, \cdots, f_N]$. We performed token enhancement by combining local tokens with global tokens: $(Q + k * G)$ and re-arranged them in the form of a 2D spatial grid which contains $\sqrt{N} = 14$ tokens each row and column (Figure 5). After that, the Global Average Pooling (GAP) was performed on the 2D grid, and finally, we divided the 2D spatial grid of tokens into N regions. The final feature vector L obtained after the GAP process can be defined as follows:

$$L_i = \frac{1}{N_R} \sum_{j=i*N_R+1}^{(i+1)*N_R} \frac{Q_j + k * G}{1 + k} \qquad i = 0, 1, \cdots, N_C - 1 \tag{7}$$

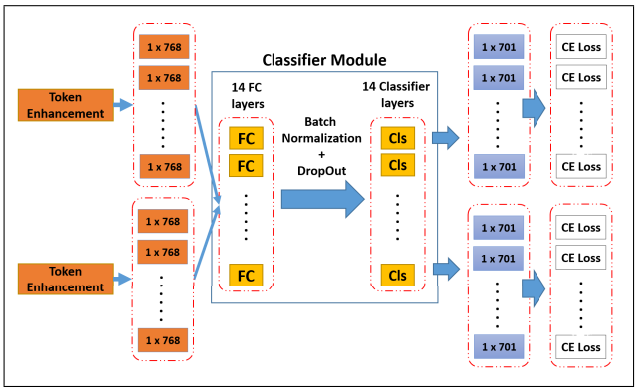


FIGURE 6. The classifier module.

where Q_i, G are local and global tokens; N_R and N_C are the total number of patches per row and column (which is $\sqrt{N}$), respectively; k is the hyperparameter which indicates the importance of global token G in the enhanced combined token. The value of k is manually set before the training (see Section IV-D2).

The final results of this process are 14 feature vectors; each has a size of 1×768 .

D. CLASSIFIER MODULE

Finally, the feature vectors from the token enhancement process are passed into the Classifier Module. The primary purpose of the Classifier module is to predict the class (geo-tag) of each image based on multiple feature vectors. The Classifier module (described in Figure 6) has $\sqrt{N}$ number of FC layers (which is 14 in this study). Each Fully-Connected (FC) layer has an input size of 768 and an output size of 512. All the final results are transferred to 14 Classifier layers (Cls), each with an input size of 512 and an output size of 701 (the number of classes).

E. LOSS FUNCTION

For the training loss, we optimize the network by calculating the Cross-Entropy loss (CE loss) on each branch of the network:

$$L_{CE}(p, y) = \begin{cases} -\log(p), & y = 1 \\ -\log(1 - p), & y = 0 \end{cases} \tag{8}$$

where p stands for the prediction result, and y stands for the class's ground-truth label (geo-tag). The final loss can be formulated as follows:

$$L_{final} = \sum_{i=0}^{N} L_{UAV}^i + L_{Satellite}^i \tag{9}$$

where L_{UAV} and $L_{Satellite}$ are the total CE losses calculated in each branch, and N is the number of feature parts (which is 14 in this work).

IV. EXPERIMENTS AND DISCUSSIONS

In Section IV, we first explain the details of the University-1652 dataset (Section IV-A) and our implementation details

TABLE 1. Detail information of the training dataset.

University-1652			
Views	Images	Classes	Universities
UAV	37,854	701	33
Satellite	701	701	33

TABLE 2. Detail information of the testing dataset.

University-1652			
Views	Images	Classes	Universities
Query (UAV)	37,854	701	33
Gallery (Satellite)	951	951	39
Query (Satellite)	701	701	33
Gallery (UAV)	51,355	951	39

(Section IV-B). After that, we introduce several experiment results in comparison with the state-of-the-art methods (Section IV-C) and demonstrate some ablation studies to get a better understanding of the proposed method (Section IV-D).

A. DATASET DETAILS

Regarding the dataset, we use the University-1652 dataset released by Zheng et al. [11], as it is the only benchmark dataset in this field that acquires both satellite-view and UAV-view images, which helps solve cross-view geo-localization for UAV navigation. This dataset contains 1,652 geographic targets from 72 universities all over the world. Each target contains three views: satellite, UAV, and street view. To reduce the high cost of collecting images by UAVs, Zheng et al. collected all UAV-view and street-view images by a 3D engine named Google Earth, while Google Map captured satellite-view images. All images in the dataset have geo-tags as their class labels. A simulated camera view controlled the view of UAVs in Google Earth, and the view height ranges from 256 to 121.5m. Each target consisted of 1 satellite-view image, 54 UAV-view images, and a few street-view images. The dataset was split into the training set and testing set. **Notice that there are no overlapped classes between two sets, which means that the images in the testing set are entirely different from the training set, and the trained model has never seen the images during the training**. The captured images have an original size of 512 × 512. The data distribution details in each set are described in Table 1 and Table 2. Samples of UAV-view images and Satellite-view images in the dataset were demonstrated in Figure 7.

B. IMPLEMENTATION DETAILS

For training, we resize all the input images to 224 224 to match ViT's original, as we want to utilize the pre-trained weights on ImageNet of ViT. We apply three types of ViT backbone: small scale-ViT, which has 8 Transformer blocks, normal-scale ViT (ViT-B), which has 12 Transformer blocks and large-scale ViT (ViT-L), which has 24 Transformer blocks. These backbones were pre-trained on ImageNet-21K [53]. The proposed model was trained over 120 epochs with a batch size of 16. We chose Stochastic Gradient Descent with Momentum (SGDM) and used a momentum of 0.9 with a weight decay of 5×10^{-4} as the optimizer for the training phase. The initial learning rate is 10^{-4} for backbone layers and 10^{-3} for other layers. All the programs were executed on Nvidia Titan XP GPU using the Pytorch framework.

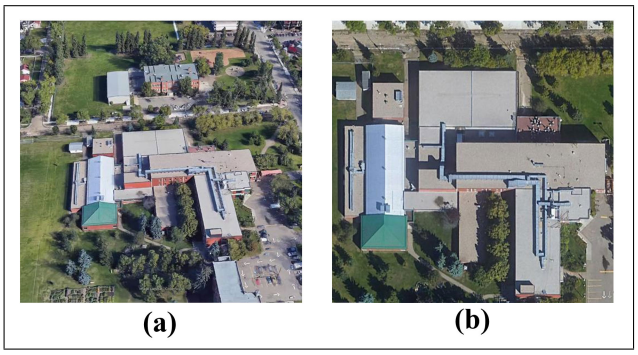


FIGURE 7. Samples of UAV-view image (a) and Satellite-view image (b) in University-1652 dataset.

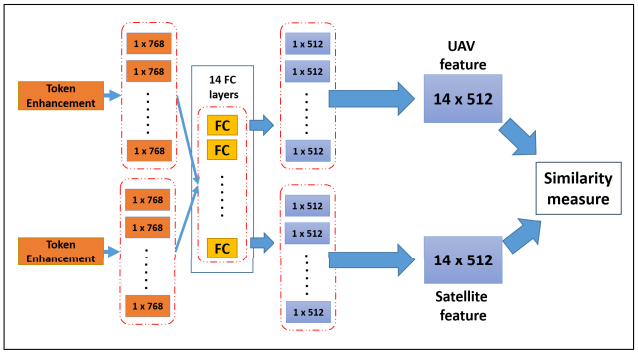


FIGURE 8. The testing stage.

For the testing stage, 14 Classifier layers (Cls) in the Classifier Module are removed, and the distance is calculated using output feature vectors extracted from each branch. As described in Figure 8, 14 feature vectors (1 × 512) are concatenated into one feature vector (14 × 512). Euclidean distance was applied to compute the similarity between feature vectors from the query and gallery:

$$D_{Euclidean} = ||F_{UAV} - F_{Satellite}||_2 \tag{10}$$

where F is the concatenated feature vector.

In this study, we evaluated our models on two tasks: UAV Satellite and Satellite UAV. We used two common measurements in cross-view geo-localization: Recall@K and Average Precision (AP).Recall@K is computed by calculating the ratio of the true-matched image in the top-K results of the ranking list. On the other hand, AP is a popular metric in measuring the precision of a retrieval system, which measures average retrieval performance with multiple ground truths. The higher Recall@K and AP, the better the model performs. We also measure the inference time of each model in the testing phase.

TABLE 3. Comparisons with state-of-the-art methods on University-1652. ViT-B represents the normal-scale ViT, ViT-S represents the small-scale ViT, and ViT-L represents the large-scale ViT. Swin-Tiny represents the small-scale Swin Transformer [46]. Inference time is measured compared to LPN [35]. The best accuracy is highlighted in **bold**.

Method	Backbone	Resolution	Inference Time	Task			
				UAV → Satellite		Satellite → UAV	
				R@1	AP	R@1	AP
Weighted Soft Margin Triplet Loss [14]	VGG16	256 × 256	1.39×	58.23	62.91	74.47	59.45
Instance Loss [33]	ResNet-50	256 × 256	-	58.49	63.13	71.18	58.74
LCM [34]	ResNet-50	256 × 256	-	66.65	70.82	79.89	65.38
LPN [35]	ResNet-50	256 × 256	1.00×	74.16	77.39	85.16	73.68
LPN [35]	ResNet-101	256 × 256	1.51×	76.13	79.29	85.45	75.45
MBSA [36]	ResNet-50	256 × 256	1.05×	82.33	82.06	91.01	82.28
SGM [24]	Swin-Tiny	224 × 224	-	79.59	82.50	87.73	79.59
SGM [24]	Swin-Tiny	256 × 256	1.04×	82.14	84.72	88.16	81.81
FSRA [25]	Vit-S	224 × 224	1.21×	80.81	83.65	87.73	80.02
FSRA [25]	Vit-S	256 × 256	1.21×	84.51	86.71	88.45	83.37
Ours	Vit-S	224 × 224	0.89×	83.88	86.25	90.87	83.65
Ours	Vit-B	224 × 224	1.28×	87.33	89.28	90.16	86.93
Ours	Vit-L	224 × 224	1.50×	**88.18**	**89.99**	**91.30**	**87.44**

TABLE 4. Ablation study on the influence of global and local tokens with different backbones. ViT-B represents the normal-scale ViT, ViT-S represents the small-scale ViT, and ViT-L represents the large-scale ViT. Swin-B represents the small-scale Swin Transformer. The best accuracy is highlighted in **bold**.

Backbone	Tokens	Number of classifiers	Task			
			UAV → Satellite		Satellite → UAV	
			R@1	AP	R@1	AP
ViT-S	Local	14	77.61	80.77	84.74	77.98
Vit-S	Local + Global	14	83.88	86.25	90.87	83.65
ViT-B	Local	14	81.19	83.99	88.30	81.72
ViT-B	Local + Global	14	87.33	89.28	90.16	86.93
ViT-L	Local	14	85.11	87.29	90.30	85.11
Vit-L	Local + Global	14	**88.18**	**89.99**	**91.30**	**87.44**
Swin-S	Local	7	79.09	82.06	85.59	78.27
Swin-S	Local + Global	7	79.52	82.45	87.02	79.09
Swin-B	Local	7	83.97	86.41	88.45	83.61
Swin-B	Local + Global	7	83.88	86.25	90.87	83.65
Swin-L	Local	7	84.72	87.04	90.30	84.74
Swin-L	Local + Global	7	84.38	86.68	90.30	83.81

C. COMPARISONS WITH THE STATE-OF-THE-ART

In Table 3, we demonstrated the performance of our proposed model compared with other state-of-the-art works. Our proposed model reached 87.33% of R@1 and 89.28% AP on UAV Satellite, 90.16% R@1 accuracy, and 86.93% AP on Satellite UAV data, which significantly exceeded the state-of-the-art MBSA [36] by a large margin of about 5% on the UAV Satellite task, and surpassed all other ResNet-50 based methods. This result confirmed that Transformer-based structures can achieve the same performance as CNN-based models and perform better with a proper feature enhancement strategy. Compared with other Transformer-based methods, our proposed model surpassed SGM [24] and FSRA [25] by nearly 5% and 3%, respectively. Additionally, with the same Transformer backbone (ViT-S) and input image resolution (224 × 224), our proposed model still achieved better results than FSRA. Especially, our proposed model with the large-scale ViT backbone (ViT-L) achieved the highest accuracy of R@1 and AP on the UAV Satellite task: 88.18% and 89.99% (nearly 4% accuracy improvement compared to all the existing state-of-the-art methods). It should be noted that the proposed model did not see the classes of testing set during the training, but the proposed model still achieved high matching accuracy. Furthermore, the inference time of the proposed model (ViT-S backbone) is only 0.89× of the ResNet-50-based model, which is faster than most existing models but still reaches competitive results.

D. ABLATION STUDIES

1) ABLATION ON THE EFFECT OF TOKEN ENHANCEMENT STRATEGY IN DIFFERENT TRANSFORMER-BASED MODELS

To further understand the influence of global and local tokens in Transformer-based models, we evaluated the performance of models that used different Transformer-based backbones and different token combinations. Here we performed several experiments on different scales of ViT (small, base, and large) that have different numbers of Transformer blocks (8, 12, and 24, respectively), and the backbone of Swin Transformer [46], which is a novel Transformer-based model that exceeded ViT on the ImageNet classification task.

In Swin-based models, the output size of the feature vector has $N = 49$. Thus, the number of classifiers in these models is changed to $\sqrt{N} = 7$. As described in Table 4, the token enhancement strategy has a positive effect on all the ViT-based models in both tasks: about 5% of accuracy increase in ViT-S and ViT-B based models, and around 1 to 3% of accuracy increase in ViT-L based models. From these results, we assumed that the global token, which contains the whole input image information in ViT, can also heavily affect the final feature representation, and a 2D grid PCB-like strategy in ViT is more effective than the re-arrangement of tokens in FSRA and SGM.

On the other hand, when we applied the token enhancement on the Swin-based models, except the Swin-S-based models, the accuracy of R@1 and AP dropped on the UAV Satellite task and had about 1% accuracy increase on the Satellite UAV task. This result concluded that the token enhancement strategy did not positively affect Swin-based models. Regarding this phenomenon, we assume that the reason lies within the difference in the way the Swin Transformer model applies the self-attention mechanism. Swin Transformer is built by replacing the standard multi-head self-attention (MSA) module in a Transformer block with a module based on shifted windows. These shifted windows build hierarchical feature maps (feature maps that are merged from layer to layer), which effectively reduce the spatial dimension of the feature maps from one layer to another. However, the tokens produced from these operations do not reserve the spatial information from the original input patch as in ViT, which makes us believe that it heavily affects the 2D grid PCB-like strategy in the final process of our proposed method.

2) INVESTIGATION ON THE EFFECT OF GLOBAL TOKEN

The hyperparameter k is an essential indicator in the token enhancement strategy. It implies the importance of global tokens in the final results. By default, we deploy $k = 8$. To verify the influence of k on the accuracy of R@1 and AP, we conducted several experiments with different values of k (in the range of 0.5 to 20) with ViT-Bas backbone, and the results are described in Figure 9. For the task of UAV Satellite, the accuracy of both R@1 and AP raised when the value of k increased and reached the highest accuracy (87% and 89%) when $k = 12$, dramatically dropping after that.

On the other hand, in the task of Satellite UAV, the accuracy of R@1 and AP have an increasing tendency when k is lower than 6, and there is an enormous instability when k increases. From these results, we assume that when k is getting bigger, the information of the global classification token may overwhelm the contextual information of local tokens and, thus, affect the feature representations that are extracted in each patch. Thus, selecting k in the training phase is quite important, and more investigations on this phenomenon should be conducted in the future.

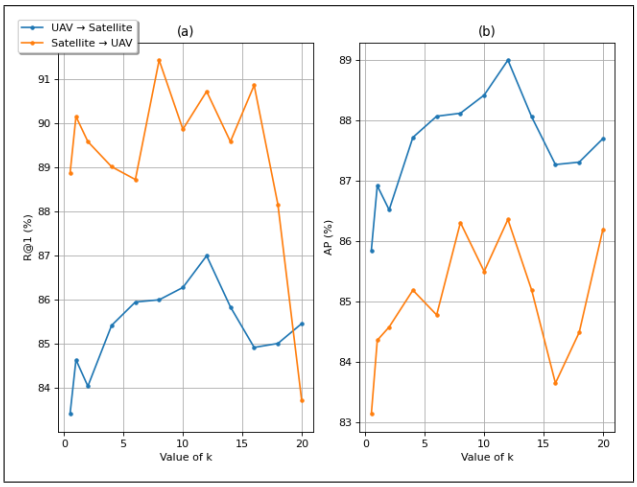


FIGURE 9. Compare the effect of the hyperparameter k on two task: UAV Satellite (blue line) and Satellite UAV (orange line). (a) Show the effect of hyperparameter k on the accuracy of Recall@1. (b) Show the effect of hyperparameter k on the accuracy of AP.

3) ABLATION OF OFFSET AND SCALE

To verify the robustness of the proposed model against offset and scale in comparison with existing methods, we performed two ablation experiments on this problem. For the first experiment, the image was shifted from 0 to 20 pixels to the right to offset the geographic target from the center view. Experiment results were compared with the results of SGM [24]. The results of the first experiment are shown in Table 5. When the offset increased from 0 to 10, on the task of UAV Satellite, the accuracy of our model dropped about 2% in accuracy (83.98% R@1 and 86.41% AP), and dropped lightly 1% when the offset increased to 20 (82.34% R@1 and 85.05% AP). Furthermore, the accuracy on the task of Satellite UAV did not change much (around 91% R@1 and 86% AP). Meanwhile, the accuracy of SGM [24] dropped dramatically on both tasks and dropped below 80% when the offset decreased to 20. This proves that our model was robust to offset and especially has consistency on the task of Satellite UAV.

In the second experiment, we tested the robustness of the model to different scales on the task of UAV Satellite. The scale of the UAV-view image in University-1652 changed dynamically during the flight. Thus, we divided the Query-UAV images into three groups: short, medium, and long, which respectively demonstrated the distances between UAV and geographic targets. Table 6 demonstrated the results of the second experiment. The model performs worse on short and long distances, achieving 83.88% and 83.74% R@1 accuracy, respectively. However, on the middle distance, the accuracy of the model surpassed the average accuracy (87.87% R@1 and 89.74% AP). The proposed model did not have a considerable margin in any level of distance. Thus, we believe the model is robust to scale and may be adapted to real-environment situations.

TABLE 5. Ablation of shifting query images during inference in comparison with existing methods.

Method	Padding pixel	Task			
		UAV → Satellite		Satellite → UAV	
		R@1	AP	R@1	AP
SGM [24]	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
Ours	0	86.00	88.12	91.44	86.31
	10	83.98	86.41	91.01	86.26
	20	82.34	85.05	91.44	86.02

TABLE 6. Ablation of using drone images with different distance to geo-graphic target in comparison with existing methods.

Distance to target	UAV → Satellite	
	R@1	AP
All	86.00	88.12
Short	83.88	86.09
Middle	87.87	89.74
Long	83.74	86.41

FIGURE 10. Visualization of qualitative result on the task of Satellite ⟶ UAV.

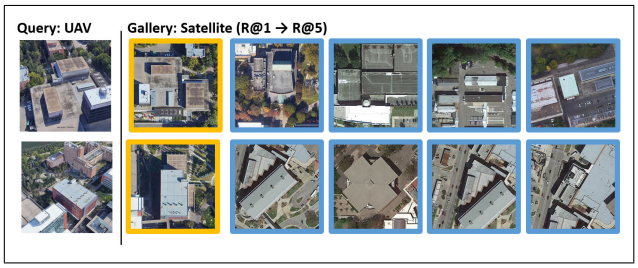


FIGURE 11. Visualization of qualitative result on the task of UAV ⟶ Satellite.

4) VISUALIZATION OF QUALITATIVE RESULT

To further verify the reliability of our proposed model, twe visualize some retrieval results of our proposed methods on two tasks (UAV Satellite and Satellite UAV). We randomly put two images of different places from UAV and Satellite query data into the proposed models. Then, we took out the top-5 most similar images from the gallery data. The true-matched images are in yellow boxes, and the false-matched images are in blue boxes. On the Satellite UAV task (Figure 10), our proposed models retrieved all the correct results from the gallery from only one query image. On the UAV Satellite task (Figure 11), even though there is only one correct answer in the gallery, the proposed model can still find the accurate satellite image.

E. DISCUSSIONS

We conducted a thorough examination of our proposed model's retrieval performance through extensive experiments in two primary tasks: UAV Satellite and Satellite UAV. In general, the Transformer-based model shows outstanding performance compared to traditional CNN-based methods. Furthermore, the participation of both global and local tokens during the retrieval phase contributed to improved image representation learning. Inference time was shortened by using a smaller ViT model (ViT-S), but the proposed model still achieved competitive results in comparison with other methods.

However, it is essential to address some certain shortcomings in our proposed method. Vision Transformer-based networks typically require high computational resources (e.g., ViT-S comprises approximately 21 million parameters, and ViT-B contains around 86 million parameters), posing several challenges for real-time mission training and deployment. Additionally, dividing extracted features into multiple parts also requires additional computational costs. Therefore, designing a lightweight Vision Transformer-based network to meet the limited computing resources on UAVs remains a significant challenge in our future research. Another issue to consider is that Vision Transformers are less robust to occlusion compared to other methods. Although ViT performs reasonably well against occlusions (ViT achieves 60% top-1 accuracy on ImageNet even when 80% of the image content is randomly occluded [54]), real-life missions may encounter situations where UAV views are obstructed by weather conditions or objects. To enhance model robustness, especially in scenarios with multiple occluding objects, we plan to further explore the correspondence between local and global tokens for representing contextual information in cross-view images.

V. CONCLUSION

With the rapid development of UAV technology, the need for autonomous control of UAVs, especially navigating UAVs without using GPS signals, is increasing rapidly. Image-based localization has been a critical solution to this problem. This paper focused on solving the cross-view image matching tasks for geo-localization. We revealed the shortcomings of existing CNN and Transformer-based methods and proposed a new Transformer-based method that utilized the local tokens and global tokens of Vision Transformer. The proposed model's performance was verified on a benchmark dataset (University-1652), and the experiment results demonstrated the outstanding performance of our model compared to the previous existing state-of-the-art methods. Notably, our method's tokenization process exhibited robust performance, promising significant potential for various applications in future Vision Transformer-related research. In the upcoming phases, we will continue to explore how to further improve the matching accuracy, especially on real UAV image data,

and seek to optimize the Transformer model for practical applications in UAVs.

REFERENCES

[1] N. A. Khan, S. N. Brohi, and N. Jhanjhi, "UAV's applications, architecture, security issues and attack scenarios: A survey," in *Proc. Intell. Comput. Innov. Data Sci. (ICTIDS)*. Singapore: Springer, Nov. 2020, pp. 753–760.

[2] G. Balamurugan, J. Valarmathi, and V. P. S. Naidu, "Survey on UAV navigation in GPS denied environments," in *Proc. Int. Conf. Signal Process., Commun., Power Embedded Syst. (SCOPES)*, Oct. 2016, pp. 198–204.

[3] A. Couturier and M. A. Akhloufi, "A review on absolute visual localization for UAV," *Robot. Auto. Syst.*, vol. 135, Jan. 2021, Art. no. 103666.

[4] K. Wang, S. Ma, J. Chen, F. Ren, and J. Lu, "Approaches, challenges, and applications for deep visual odometry: Toward complicated and emerging areas," *IEEE Trans. Cogn. Develop. Syst.*, vol. 14, no. 1, pp. 35–49, Mar. 2022.

[5] C. Chen, B. Wang, C. Xiaoxuan Lu, N. Trigoni, and A. Markham, "A survey on deep learning for localization and mapping: Towards the age of spatial machine intelligence," 2020, *arXiv:2006.12567*.

[6] L. Yu, E. Yang, B. Yang, Z. Fei, and C. Niu, "A robust learned feature-based visual odometry system for UAV pose estimation in challenging indoor environments," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–11, 2023.

[7] A. Steenbeek and F. Nex, "CNN-based dense monocular visual SLAM for real-time UAV exploration in emergency conditions," *Drones*, vol. 6, no. 3, p. 79, Mar. 2022.

[8] R. Ali, D. Kang, G. Suh, and Y.-J. Cha, "Real-time multiple damage mapping using autonomous UAV and deep faster region-based neural networks for GPS-denied structures," *Autom. Construction*, vol. 130, Oct. 2021, Art. no. 103831.

[9] 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.

[10] D. Wilson, X. Zhang, W. Sultani, and S. Wshah, "Visual and object geo-localization: A comprehensive survey," 2021, *arXiv:2112.15202*.

[11] 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.

[12] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: Speeded Up Robust Features," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2006, pp. 404–417.

[13] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. 7th IEEE Int. Conf. Comput. Vis.*, vol. 2, Sep. 1999, pp. 1150–1157.

[14] 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.

[15] 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.

[16] Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li, "Optimal feature transport for cross-view image geo-localization," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, pp. 11990–11997.

[17] Y. Tian, C. Chen, and M. Shah, "Cross-view image matching for geo-localization in urban environments," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1998–2006.

[18] S. Cai, Y. Guo, S. Khan, J. Hu, and G. Wen, "Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8390–8399.

[19] R. Rodrigues and M. Tani, "Are these from the same place? Seeing the unseen in cross-view image geo-localization," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3752–3760.

[20] D. V. Bui, M. Kubo, and H. Sato, "A part-aware attention neural network for cross-view geo-localization between UAV and satellite," *J. Robot., Netw. Artif. Life*, vol. 9, no. 3, pp. 275–284, Dec. 2022.

[21] D. V. Bui, M. Kubo, and H. Sato, "Attention-based neural network with generalized mean pooling for cross-view geo-localization between UAV and satellite," *Artif. Life Robot.*, vol. 28, pp. 560–570, Apr. 2023.

[22] H. Yang, X. Lu, and Y. Zhu, "Cross-view geo-localization with layer-to-layer transformer," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 29009–29020.

[23] S. Zhu, M. Shah, and C. Chen, "TransGeo: Transformer is all you need for cross-view image geo-localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1152–1161.

[24] J. Zhuang, X. Chen, M. Dai, W. Lan, Y. Cai, and E. Zheng, "A semantic guidance and transformer-based matching method for UAVs and satellite images for UAV geo-localization," *IEEE Access*, vol. 10, pp. 34277–34287, 2022.

[25] M. Dai, J. Hu, J. Zhuang, and E. Zheng, "A transformer-based feature segmentation and region alignment method for UAV-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4376–4389, Jul. 2022.

[26] S. Workman, R. Souvenir, and N. Jacobs, "Wide-area image geolocalization with aerial reference imagery," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3961–3969.

[27] 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. 5617–5626.

[28] K. Kavitha and B. T. Rao, "Evaluation of distance measures for feature based image registration using AlexNet," 2019, *arXiv:1907.12921*.

[29] 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*, Nov. 2011, pp. 1125–1128.

[30] 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.

[31] S. Workman and N. Jacobs, "On the location dependence of convolutional neural network features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 70–78.

[32] N. N. Vo and J. Hays, "Localizing and orienting street views using overhead imagery," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 494–509.

[33] 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.

[34] 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.

[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.

[36] 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 Sens.*, vol. 13, no. 19, p. 3979, Oct. 2021.

[37] 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. (ECCV)*, 2018, pp. 480–496.

[38] 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 worth 16 × 16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[39] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," 2020, *arXiv:2010.04159*.

[40] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "DeepViT: Towards deeper vision transformer," 2021, *arXiv:2103.11886*.

[41] 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," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 6877–6886.

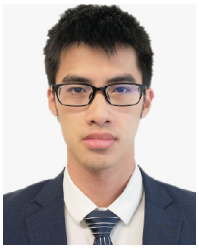
[42] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 213–229.

[43] 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. 14993–15002.

[44] Y. Lu, M. Jiang, Z. Liu, and X. Mu, "Dual-branch adaptive attention transformer for occluded person re-identification," *Image Vis. Comput.*, vol. 131, Mar. 2023, Art. no. 104633.

[45] X. Tian, J. Shao, D. Ouyang, A. Zhu, and F. Chen, "SMDT: Cross-view geo-localization with image alignment and transformer," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2022, pp. 1–6.

[46] 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. 9992–10002.
[47] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 22–31.
[48] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, and H. Xue, "Towards robust vision transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12032–12041.
[49] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," 2020, *arXiv:2012.09958*.
[50] Z.-H. Jiang, Q. Hou, L. Yuan, D. Zhou, Y. Shi, X. Jin, A. Wang, and J. Feng, "All tokens matter: Token labeling for training better vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 18590–18602.
[51] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. H. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 538–547.
[52] C. Sharma, S. R. Kapil, and D. Chapman, "Person re-identification with a locally aware transformer," 2021, *arXiv:2106.03720*.
[53] T. Ridnik, E. Ben-Baruch, A. Noy, and L. Zelnik-Manor, "ImageNet-21K pretraining for the masses," 2021, *arXiv:2104.10972*.
[54] M. M. Naseer, K. Ranasinghe, S. H. Khan, M. Hayat, F. S. Khan, and M.-H. Yang, "Intriguing properties of vision transformers," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds. Red Hook, NY, USA: Curran Associates, 2021, pp. 23296–23308.



DUC VIET BUI was born in Hanoi, Vietnam, in 1995. He received the M.S. degree from the Department of Computer Science, National Defense Academy, Japan, in 2021, where he is currently pursing the Ph.D. degree with the Department of Computer Science. His research is related to different applications of computer vision in aerial robotics. His research interests include computer vision, machine learning, deep neural networks, and aerial robotics.

MASAO KUBO received the degree from the Precision Engineering Department, Hokkaido University, in 1991, and the Ph.D. degree in computer science (multi-agent system) from Hokkaido University, in 1996. He is an Associate Professor with the Department of Computer Science, National Defense Academy, Japan. He was a Research Assistant with the Chaotic Engineering Laboratory, Hokkaido University. He was also a Visiting Research Fellow of the Intelligent Autonomous Laboratory, University of the West of England, in 2005. He is an Associate Professor with the Information System Laboratory, Department of Computer Science, National Defense Academy. His research interest includes multi-agent systems.

HIROSHI SATO received the degree in physics from Keio University, Japan, and the master's and Ph.D. degrees in engineering from the Tokyo Institute of Technology, Japan. He is an Associate Professor with the Department of Computer Science, National Defense Academy, Japan. Previously, he was a Research Associate with the Department of Mathematics and Information Sciences, Osaka Prefecture University, Japan. His research interests include agent-based simulation, evolutionary computation, and artificial intelligence. He is a member of the Japanese Society for Artificial Intelligence (JSAI), the Society of Instrument and Control Engineers (SICE), and the Institute of Electronics, Information and Communication Engineers (IEICE). He was an editor of IEICE and SICE.