A Transformer and Visual Foundation Model-Based Method for Cross-View Remote Sensing Image Retrieval

Changjiang YIN, Qin YE *, Junqi LUO

College of Surveying and Geo-Informatics, Tongji University, 200092, Shanghai, China
– yeqin@tongji.edu.cn, sika@tongji.edu.cn, harden@tongji.edu.cn

Keywords: Image Retrieval, Cross-view, Feature Space, Vision Transformer, Visual Foundation Model.

Abstract

Retrieving UAV images that lack POS information with georeferenced satellite orthoimagery is challenging due to the differences in angles of views. Most existing methods rely on deep neural networks with a large number of parameters, leading to substantial time and financial investments in network training. Consequently, these methods may not be well-suited for downstream tasks that have high timeliness requirements. In this work, we propose a cross-view remote sensing image retrieval method based on transformer and visual foundation model. We investigated the potential of visual foundation model for extracting common features from cross-view images. Training is only conducted on a small, self-designed retrieval head, alleviating the burden of network training. Specifically, we designed a CVV module to optimize the features extracted from the visual foundation model, making these features more adept for cross-view image retrieval tasks. And we designed an MLP head to achieve similarity discrimination. The method is verified on a publicly available dataset containing multiple scenes. Our method shows excellent results in terms of both efficiency and accuracy on 15 sub-datasets (10 or 50 scene categories) derived from the public dataset, which holds practical value in engineering applications with streamlined scene categories and constrained computational resources. Furthermore, we initiated a comprehensive discussion and conducted ablation experiments on the network design to validate its efficacy. Additionally, we analyzed the presence of overfitting within the network and deliberated on the limitations of our study, proposing potential avenues for future enhancements.

1. Introduction

Oblique UAV images have become increasingly pivotal in applications such as urban modeling and scene understanding (Verykokou et al., 2016, Sheppard and Rahnemoonfar, 2017). Determining the geo-location of oblique UAV images accurately is the fundamental basis of these applications. When oblique UAV images lack POS information, we usually retrieve and subsequently register these images with georeferenced satellite orthoimagery. It is a cross-view remote sensing image retrieval task, and is challenging due to the substantial differences in angles of view between oblique UAV images and satellite images. Traditional handcrafted feature-based image retrieval methods struggle to capture common features between such cross-view images. Nowadays, most cross-view image retrieval methods are grounded in deep learning approaches. However, many methods aim to enhance retrieval performance by stacking learning modules. This augmentation increases the number of model parameters and poses challenges in fine-tuning tasks when the scene changes, thereby limiting the practical applicability of these methods in engineering.

Recently, Meta AI Research introduced DinoV2 (Oquab et al., 2023), a large visual foundation model that demonstrates robust generalization and zero-shot transfer capabilities in downstream tasks such as semantic segmentation and depth estimation. The latent features extracted by DinoV2 exhibit exceptionally strong common feature representation capabilities in contrast to the backbones utilized for single tasks, such as image matching. Thus DinoV2 can provide a solid foundation for initializing image features in cross-view remote sensing image retrieval, thereby facilitating accurate model regression. Furthermore, the use of DinoV2 for cross-view remote sensing image retrieval obviates the need for weight fine-tuning, thus avoiding extensive training during the fine-tuning task and reducing computational costs. The key to effectively leveraging DinoV2 lies in the design of an effective downstream task head.

Therefore, in this work, we investigate a novel cross-view remote sensing image retrieval method based on a visual foundation model DinoV2, and a transformer-based retrieval head. In summary, our contributions include:

1. We employed zero-shot transfer learning on the backbone network. We introduced the vision foundation model as the backbone for the cross-view image retrieval task and froze its weight, thus circumventing the additional cost from the backbone network training.

2. We designed a retrieval head, a small network based on transformer with only a few parameters. We not only alleviated the burden on network training but also enhanced the features from DinoV2, rendering them more adept for retrieval tasks.

3. We proposed a novel deep learning approach for cross-view image retrieval, which combines contrastive learning, supervised learning, and transfer learning, integrating these latest techniques in the field of deep learning.

2. Related Work

2.1 Cross-View Remote Sensing Image Retrieval

Cross-view image retrieval is widely used for rough positioning of query images. Most methods for cross-view image retrieval follow a standard data processing pipeline. First, features of

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-1-2024-821-2024 | © Author(s) 2024. CC BY 4.0 License.

query and gallery images are extracted individually, then the similarities between these images are measured, and finally, the gallery image with the highest similarity score is selected as the retrieval result. Some researchers employed handcrafted features for cross-view image characterization. Cheng et al. (2018) used SIFT descriptors for retrieval between cross-view ground images. Luo and Ye (2023) designed the SDS (Segments Direction Statistics) feature pattern, and used it in the oblique-view UAV image-based retrieval task. Owing to the significant advantages of neural networks in feature extraction and the continual development of cross-view image datasets like University-1652 (Zheng et al., 2020), cross-view image retrieval research has predominantly relied on deep learning methodologies over the past decade.

Zemene et al. (2019) designed a retrieval method for querying in a city-wide reference image database with known absolute coordinates, thereby determining the geo-location of the query image. Similarly, Rodrigues and Tani (2022) performed retrieval between ground images and a large geotagged aerial image database. Some methods enhance the retrieval capabilities of the network by improving training strategies or modifying the framework of the model. Zhang et al. (2022) proposed a deep neural network that introduced a spatial scale attention mechanism for cross-view image feature extraction, strengthening the scene spatial information representation. Lin et al. (2022) presented a feature learning approach based on joint learning, leveraging a single network to acquire discriminative features. They also introduced a key point detection model to emulate human visual perception, thereby enhancing the feature's capability to represent key areas. Zeng et al. (2022) designed a peer learning-based parallel retrieval method incorporating two siamese networks. They utilized UAV images as intermediaries between ground images and satellite images to facilitate retrieval between the two views. Hu et al. (2018) developed a global feature generation module to further optimize the local features extracted by the backbone network. Additionally, they introduced a weighted soft margin ranking loss to accelerate model convergence.

Some recent studies have opted for transformer-based backbone networks instead of CNNs. For instance, the FSRA (Dai et al., 2021) automatically divided the original image into multiple regions based on the heat distribution of the feature map, and achieved feature alignment based on region consistency. Zhuang et al. (2022) introduced semantic constraints based on FSRA to enhance the effectiveness of feature alignment. However, a limitation shared by all the above methods is that the training process still involves the backbone network, resulting in increased computational and time overhead for the retrieval task. Therefore, in this study, we introduce an approach by incorporating a visual foundation model with robust generalization capabilities as the backbone network of our model.

2.2 Visual Foundation Model

The effectiveness of a neural network lies in the proper initialization of its parameters. In the field of deep learning, this initialization process necessitates a substantial amount of high-quality training data. However, many downstream tasks lack access to such data. Therefore, a common approach for most tasks is to fine-tune foundation models that have been pre-trained on large datasets. Vision foundation models are widely used for transfer learning (Zhou et al., 2023). Initially, these models referred to pre-trained weights of backbone networks obtained by training CNN networks (for example, ResNet) on general and labeled datasets (including ImageNet). These pre-trained weights were then transferred to downstream tasks to expedite convergence. However, due to the expense of data annotation, the size of these datasets is limited, and the network's generalization performance cannot be guaranteed. Most subsequent research has concentrated on semi-supervised learning or self-supervised learning methods, as weakly labeled or unlabeled data is generally more accessible.

The visual foundation model typically consists of an encoder and a decoder. In transfer learning, fine-tuning the encoder part is generally focused, while the task-specific heads are connected to various downstream tasks. CNNs were previously utilized to construct visual foundation models. Context (Doersch et al., 2015) is a self-supervised learning method that learns the contextual information in the image through random sampling patches, thus enhancing the semantic attributes of features. The vision transformer (ViT) has emerged as a prominent research focus in recent years due to its capability to achieve superior training results on large datasets. BEiT (Bao et al., 2021) introduced patch random masking based on the classic ViT, forcing the network to strengthen the representation ability of latent features. SAM (Kirillov et al., 2023) is a visual foundation model for segmentation tasks. It has achieved extremely robust semantic segmentation performance by training on the SA-1B dataset which contains 1 billion masks. Although visual foundation models demonstrate strong generalization capabilities and have been widely employed in various downstream tasks, their application in the field of cross-view image retrieval remains unexplored. DinoV2 (Oquab et al., 2023) presently stands as one of the most widely adopted visual foundation models for downstream tasks. It has acquired robust generalization capabilities through training on the extensive LVD-142M dataset and is capable of achieving zero-shot transfer. Therefore, in this work, we designed a cross-view retrieval method based on DinoV2.

3. Methodology

The overall framework of our model is shown in Fig. 1. We designed a siamese network, comprising three parts: 1) The visual basic model functions as the backbone network, extracting local and global features from UAV and satellite images. 2) The cross-view ViT (CVV) module serves as the feature adaption module, enhancing the features extracted from the backbone to align with the requirements of the cross-view image retrieval task. 3) The classification head receives two sets of image features and is responsible for identifying geospatial relations. Since we use both supervised and contrastive learning, we also illustrate our loss function design.

3.1 Visual Foundation Model Backbone

DinoV2 consists of both an encoder and a decoder, and performs discriminative self-supervised pre-training on the large LVD-142M dataset, achieving robust zero-shot transfer generalization capabilities. In this work, we transfer the DinoV2 encoder as the backbone network. The original DinoV2 encoder is a ViT model containing 1 billion adjustable parameters, which places high demands on the hardware even for the inference process. Therefore, the unsupervised distillation method is employed in DinoV2, where the original model serves as a teacher model and is compressed into three student models of varying sizes to accommodate different downstream task application scenarios. Since the patch size is set to 14 in

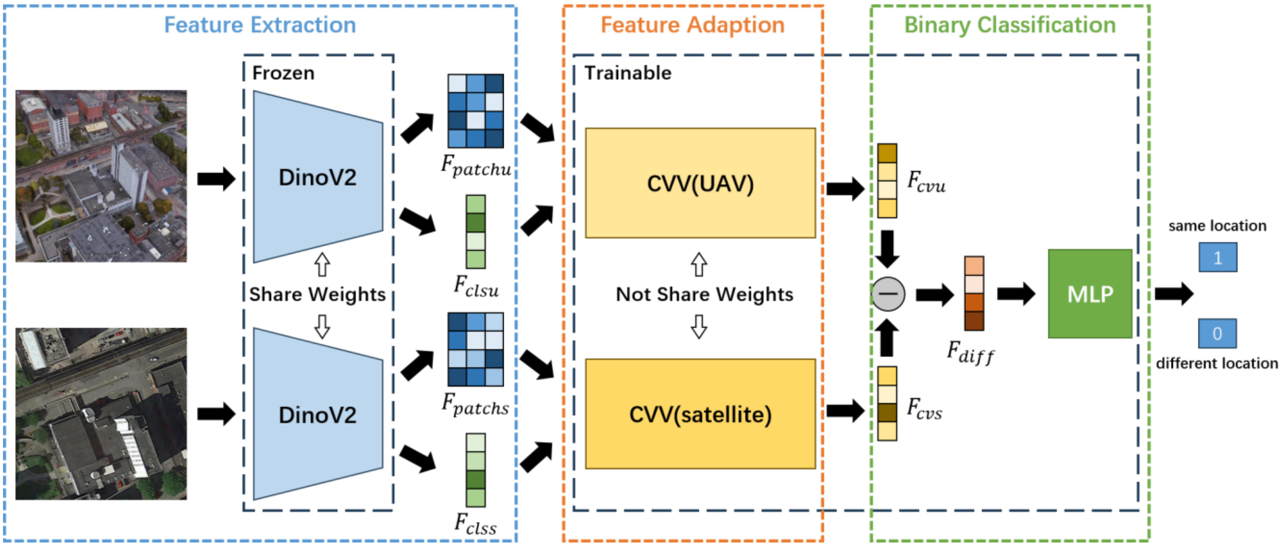

Figure 1. Research framework.

DinoV2, these four models are called ViT-G/14, ViT-L/14, ViT-B/14, and ViT-S/14 respectively.

The generalization performance of the student models is similar to that of the teacher model, despite the significant reduction in the number of parameters. Considering the limited hardware resources in engineering and the real-time requirements of cross-view retrieval tasks, we opt to utilize ViT-S/14 (21m params) as our backbone network. Ibrahimovic et al. (2023) asserted that in ViT, the number of patches significantly impacts the model's performance in downstream tasks. Excessive patches can lead to increased computational costs and potential overfitting on the training set. Therefore, we ultimately decide to set the number of patches to 16×16. Since the patch size is 14, our image input size is determined to be 224×224. The author of University-1652 noted that the input size has few impacts on cross-view image retrieval performance. The difference in Recall@1 performance between the 224×224 input size setting and the best setting is less than 3%. Therefore, it is believed that setting the input size to 224×224 will not significantly affect model performance. However, it can substantially reduce the computational burden during model training and inference. To inherit all the prior knowledge from DinoV2 and achieve a comprehensive representation of cross-view images, we simultaneously extracted both global features (size of 1×384) and local features (size of 256×384) of the image, as illustrated in formula (1):

$$F_{cls}, F_{patch} = Enc_{vits14}(X), \tag{1}$$

where F_{cls}, F_{patch} = global features and local features
Enc_{vits14} = ViT-S/14 backbone from DinoV2
X = input image

3.2 Cross-view Feature Adaptation Module

Although DinoV2 has demonstrated strong zero-sample transfer capabilities, conducting downstream tasks directly based on the global features extracted by the DinoV2 encoder presents challenges (Lu et al., 2019). Therefore, optimization of the latent features obtained through DinoV2 encoding is necessary to adapt to cross-view image retrieval tasks. In certain research based on visual foundation models, a feasible approach involves adding a feature adaptation module after the backbone network to align features with downstream tasks (Houlsby et al., 2019). We adopt this feature optimization technique. Given the exceptionally strong generalization abilities of features obtained from DinoV2, we freeze the weights of the backbone network during feature optimization. Consequently, we preserve the zero-shot transfer characteristics of DinoV2. Since the encoder of DinoV2 is constructed using ViT architecture, to ensure consistency in feature representation, our feature adaptation module is also constructed based on ViT, named cross-view ViT (CVV), as shown in Fig. 2.

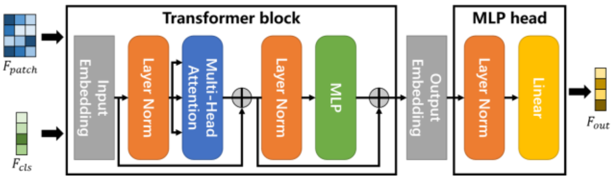

Figure 2. Architecture of CVV

Our design draws inspiration from existing work utilizing visual foundation models (Cheng et al., 2022), comprising a ViT module and a MLP head. This configuration facilitates the optimization of both global and local features extracted from DinoV2, resulting in the generation of new global features. The ViT module is composed of only one transformer block. This decision is informed by the fact that the latent feature extracted by DinoV2 already possesses strong generalization capabilities. Therefore, a simplified network suffices to optimize the feature space, aligning it with the requirements of the cross-view retrieval task. Furthermore, this design ensures the minimization of trainable parameters in the network, thereby effectively reducing computational and time costs in the subsequent training process. A classic ViT network initially divides the input image into image patches, and then converts each image patch into an embedding vector through a linear transformation. In CVV, we directly utilize global features and local features from the backbone network as embedding vectors. These embedding vectors, along with position codes, are concatenated into a sequence to generate a new embedding, which is subsequently input into the transformer block. Then, the embedding vectors containing global features, local features and position codes interact and undergo nonlinear transformations through self-attention mechanisms and feed-forward neural networks. This process facilitates the capture of global contextual information. Finally, the embedding vector at a specific position in the output sequence is sent to the MLP head for a nonlinear change to obtain optimized new features. The above process can be written as a formula (2).

$$\begin{aligned}
&E^{'} = \left[F_{cls}, F_{patch} \right] + E_p \\
&E_a = LayerNorm(E^{'} + MultiHead(E^{'})) \\
&E_f = LayerNorm(E_a + FFN(E_a)) \\
&e_{cls} = E_f \left[i \right] \\
&F_{out} = MLPhead(e_{cls})
\end{aligned} , \quad (2)$$

where E_p = position codes
$E^{'}$ = input embedding sequence
e_{cls} = output embedding of transformer block
F_{out} = optimized global feature

In cross-view image retrieval, as it essentially boils down to a classification task, our primary objective is to capture global features from the CVV module. Given variations in angles of views, feature representations of satellite and UAV images may differ, thus the CVV module in both the satellite and UAV branches don't share weights.

3.3 Classification Head

After obtaining the global features corresponding to satellite and UAV images respectively, it is necessary to determine the geospatial relation of the two images based on these features to complete the retrieval task. In many studies, discrimination based on cosine similarity or Euclidean distance serves as a prevalent method, where the group of images with the highest score is selected as the retrieval result. These methods still require additional similarity evaluation operations after feature extraction based on neural networks, thus we design an end-to-end method. We simplify the cross-view image retrieval task into a classification task and employ an MLP head to perform similarity evaluation. As the scenes in the training data may differ from those encountered during actual usage encoding the geographical location of the scene as a classification label proves challenging. Consequently, implementing a multi-classification-based cross-view retrieval model becomes difficult. Inspired by (Zhou et al., 2023), we designed the network as a binary classification model, as shown in Fig. 1. The positive category signifies that the UAV image and the satellite image describe the same geographical space, while the negative category indicates that they were captured at distinct geospatial locations. To achieve feature interaction, we adopt feature subtraction between the features extracted from the UAV and satellite images, thus obtaining the new discriminative feature that represents the differences between the two views. The new feature is then fed into the MLP head for spatial relationship discrimination. We utilize a linear layer to compress the feature into a one-dimensional representation, followed by normalization using the sigmoid to express the probability of a positive class. This process can be expressed as formula (3).

$$\begin{aligned}
&F_{diff} = F_{cvu} - F_{cvs} \\
&F_n = LayerNorm(F_{diff}) \\
&S_p = sigmoid(Linear(F_n))
\end{aligned} , \quad (3)$$

where F_{cvu}, F_{cvs} = global feature of UAV and satellite
F_{diff} = discriminative feature
S_p = similarity score

Due to potential Internal Covariate Shift issues, the distribution of varying discriminative features may be inconsistent, leading to difficulty in retrieval. In order to enhance the robustness and accelerate convergence in the training process, we incorporated layer norm for feature normalization before inputting the linear layer, following the design of the MLP module in CVV.

3.4 Loss Function

Contrastive learning is a widely applied deep learning technique utilized in siamese networks. It enables the acquisition of a more discriminative feature representation method to assist downstream tasks by learning the consistencies between positive samples and mining the differences among negative samples. However, recent studies predominantly integrate contrastive learning with supervised learning. This is attributed to the availability of geospatial location labels in cross-view image retrieval datasets used for training, and numerous research have shown the performance enhancements achievable through supervised learning. Hence, our method also combines contrastive learning and supervised learning, and sets corresponding loss functions respectively. Contrastive loss aims to minimize the Euclidean distance between positive samples and optimize the Euclidean distance between negative samples to a fixed value, thereby augmenting the differentiation between positive and negative samples. Since we have designed a binary classification head, we adopt the classic Binary Cross-Entropy Loss (BCELoss) as the loss function for supervised learning. Our final loss function incorporates both contrastive loss and BCEloss. Balancing model performance and generalization, we set the weight coefficients of both losses to 1. The loss function is shown in formula (4).

$$\begin{aligned}
&L_C = \frac{1}{2N} \sum_{i=1}^{N} \left[y_i d_i^2 + (1 - y_i) \max(margin - d_i, 0)^2 \right] \\
&L_{BCE} = -\frac{1}{N} \sum_{i=1}^{N} \left[y_i log(p_i) + (1 - y_i) log(1 - p_i) \right] \\
&L = L_C + L_{BCE}
\end{aligned} , \quad (4)$$

where N = number of training samples
y_i = label of i th sample, 1 represents positive
d_i = Euclidean distance of i th sample
p_i = predicted probability of i th sample
$margin$ = distance threshold hyperparameter

The contrastive loss is employed to optimize the feature space of UAV and satellite global features obtained by the CVV module, thereby making these features more discriminative to better adapt to retrieval tasks. The BCELoss optimizes the similarity score output by the classification head. As the contrastive loss involves the setting of a hyperparameter margin, given the size of the global feature which output by the CVV module is 1*384, and considering that the feature is also layer normed, we set the margin to 10.

4. Experiments

4.1 Experimental Setup

4.1.1 Dataset: We chose the University-1652 dataset (Zheng et al., 2020), which is widely recognized as a benchmark dataset for cross-view image retrieval tasks. The University-1652 dataset comprises images captured from three different viewpoints: UAV, satellite, and ground, each tagged with a four-digit geospatial label. The UAV images are simulated using Google Earth and collected from various scenes. Our experiments only selected images from UAVs and satellites, as shown in Fig. 3. In the dataset configuration, both the training set and the test set comprise images of 701 distinct buildings. Additionally, images of another 250 buildings are included in the test set as interference. For each building, there are 54 UAV images from various view angles and one corresponding satellite image. Some UAV images exhibit large oblique angles, which poses a challenge. Since our model employs supervised learning and utilizes supervision types of positive and negative, we transformed the two original geospatial labels in a set of images into a discriminative label of 1 or 0. Here, 1 represents positive samples and 0 represents negative samples. We randomly selected 15 sub-datasets from the test set to evaluate our method, forming 2 evaluation sets respectively. 10 sub-datasets (evaluation set1) each contain 10 scenes, while the remaining 5 sub-datasets (evaluation set2) consist of 50 scenes each. Evaluation set2 contains more scene categories than evaluation set1, and the composition is more complex. The scenes in different sub-datasets do not overlap. We devised this setting with a focus on engineering application scenarios. Some applications entail fewer scene categories, while others involve relatively more. Our setting considered both application scenarios (10 and 50 scenes), thereby ensuring a comprehensive evaluation of our method.

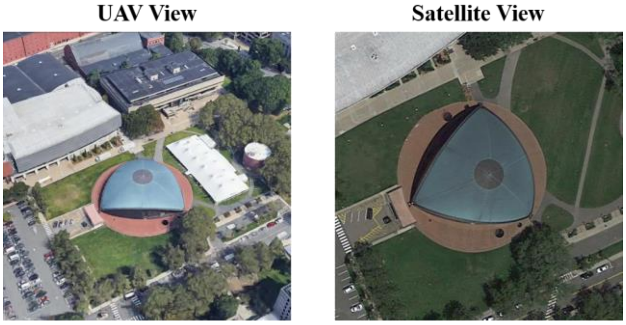

Figure 3. An example of University-1652

4.1.2 Evaluation Protocol: Recall@k is the most widely used evaluation metric in cross-view image retrieval tasks. Compared to Recall@k, Average Precision (AP) is a more comprehensive evaluation metric that considers both Precision and Recall across various thresholds. Therefore, we employed Recall@k and AP for evaluation. Recall@k indicates that at least one of the top k retrieval results, ranked by similarity score, corresponds to a positive sample. AP is computed by calculating the area under the Precision-Recall curve. We focused on Recall@1 and Recall@3 in this work.

4.1.3 Implementation Details: We balanced model performance and training cost, and the size of all input images was adjusted to 224×224. Given that the original image size is 512×512, to preserve the details, we utilized cubic interpolation. We normalized the input to minimize the disparity between samples, enhance model stability and generalization, and expedite convergence. Additionally, to further enhance the model's generalization capability, we applied data augmentation on both UAV and satellite images in the training set, including random cropping, padding, horizontal and vertical flipping, etc. Since we only train the CVV module and classification head, we froze the backbone network from DinoV2. To minimize additional computational cost, we utilized ViT-S/14 to pre-extract global and local features of the image and store them in memory. We configured the batch size to 256 and utilized the AdamW optimizer for training. The weight decay was set to 0.0005, while the initial learning rate was set to 0.001. We employed a learning rate scheduler that reduced the learning rate to 10% of its previous value every 50 epochs, and the training process spanned a total of 200 epochs.

4.2 Experiments Results

We initially analyzed the evaluation set 1. As depicted in Tab. 1, our method attained a Recall@1 exceeding 80% across the ten sub-datasets within evaluation set 1. Notably, five of these sub-datasets exhibited Recall@1 values surpassing 90%, with the average Recall@1 across all ten sub-datasets reaching 89.29%. Regarding the AP metric, all sub-datasets demonstrated an AP exceeding 80%, with over half of them surpassing 90%. The average AP across all sub-datasets reached 89.86%. Thus, our method achieved superior retrieval performance for cross-view image retrieval across ten different scenes. Our method demonstrated robustness, with the highest sub-dataset Recall@1 reaching 94.52%, and even the lowest sub-dataset Recall@1 reaching 80.89%. The variance in Recall@1 and AP metrics across different sub-datasets can be attributed to variations in the distribution of each subset.

Dataset	Recall@1	AP
Sub-dataset1	92.87	94.35
Sub-dataset2	93.28	93.02
Sub-dataset3	87.73	88.93
Sub-dataset4	84.23	80.16
Sub-dataset5	92.66	90.60
Sub-dataset6	91.22	89.24
Sub-dataset7	86.49	91.72
Sub-dataset8	94.52	94.85
Sub-dataset9	88.91	91.29
Sub-dataset10	80.98	84.40
Average	89.29	89.86

Table 1. Results of evaluation set1

In evaluation set2, only one sub-dataset's Recall@1 fell below 60%, with sub-dataset14 achieving a Recall@1 of 74.14%, and the average Recall@1 across the five sub-datasets reaching 64.50%. For the AP metric, all sub-datasets exhibited AP values exceeding 60%, with the average AP across all sub-datasets reaching 68.67%. Even more surprising is the Recall@3 metric, as all sub-datasets exhibited Recall@3 values surpassing 80%, with the highest achieving an AP of 92.24% in sub-dataset14. We specifically evaluated Recall@3 in this context due to the complex scene composition in evaluation set2. A high

Recall@3 metric addresses the requirements of engineering applications effectively. The presence of a greater number of scene categories within each sub-dataset in evaluation set2, accentuates the disparity in sample distribution among the sub-datasets. This discrepancy is also evident in the Recall@1 and AP. The results reflected the randomness of our experiments and verified the robustness and effectiveness of our method.

Dataset	Recall@1	Recall@3	AP
Sub-dataset11	58.30	84.39	61.22
Sub-dataset12	60.29	80.83	68.84
Sub-dataset13	60.58	85.68	68.12
Sub-dataset14	74.14	92.24	71.86
Sub-dataset15	69.21	88.57	73.32
Average	64.50	86.34	68.67

Table 2. Results of evaluation set2

Overall, our method showed excellent retrieval performance, effectively meeting the requirements of engineering with scene categories less than 50. We also present the performance of several classic methods for cross-view image retrieval tasks on the University-1652 dataset in Tab. 3. The direct comparison of our method with these classic methods may not be appropriate due to differences in the datasets used for evaluation. Among the listed methods, FSRA's demonstrates significantly superior performance compared to others. This is attributed to its utilization of a large backbone network (21 million parameters) and the adoption of ViT instead of CNN. These factors enable FSRA to better capture global features, a crucial aspect in cross-view image retrieval tasks. However, the number of trainable parameters of these methods exceeds 20m, while our method comprises only 2.8m parameters. This significant reduction in parameters greatly diminishes both the training and the inference time of the network. Therefore, our method possesses distinct advantages in engineering.

Method	Recall@1	AP
Zheng et al.	52.39	57.44
LCM	66.65	70.82
LPN	75.93	79.14
FSRA	84.51	86.71

Table 3. Classic methods performance on University-1652

5. Discussions

5.1 Enhanced feature discrimination with introduced CVV

The CVV module stands out as the cornerstone of our work, as it plays a pivotal role in inheriting latent features from the DinoV2 backbone network and facilitating feature adaptation to suit cross-view image retrieval tasks. As we calculate the similarity between satellite and UAV images and ascertain geospatial relations based on the global features extracted by the CVV module, the effectiveness of the CVV module profoundly influences the performance of the network in cross-view image retrieval tasks. To evaluate the effectiveness of CVV in improving feature discrimination is necessary. Here we compile statistics on the distribution of Euclidean distances between the global features of positive and negative samples. We randomly selected 10,000 sets of cross-view images from the test set, comprising 5,000 positive examples and 5,000 negative examples. To validate the enhancement achieved by CVV adaptation in the feature space, we assessed the distribution of global features directly from DinoV2 and the global features after CVV adaptation. Results are shown in Fig. 4 and Fig. 5.

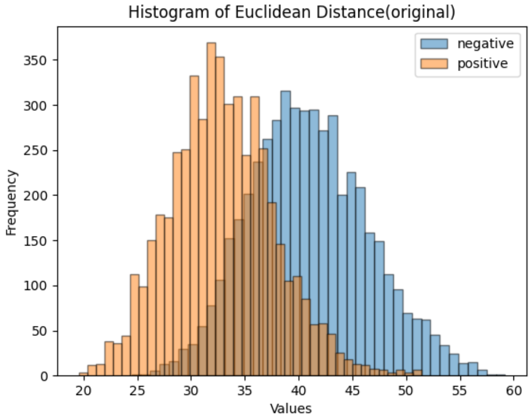

Figure 4. Distribution of Euclidean distance between global features from DinoV2 backbone

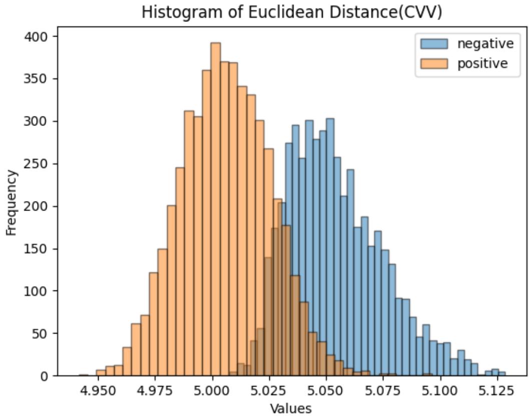

Figure 5. Distribution of Euclidean distance between global features after CVV's adaption

Although the DinoV2 backbone network underwent pre-training using entirely different datasets and deep learning techniques, an examination of the positive and negative class distributions reveals distinct patterns between the two distributions. This observation suggests that DinoV2 already possesses certain discriminative capabilities in cross-view image retrieval tasks. This result also confirms that it is reasonable to choose DinoV2 as the backbone network of our cross-view image retrieval model. Nevertheless, the global features extracted by DinoV2 still exhibit a big overlap in the spatial distribution between positive and negative classes. This observation elucidates the challenge of conducting direct retrieval solely based on the global features derived from DinoV2. Following the application of the CVV module for feature adaptation, the overlap between the two distributions notably diminishes, and two peaks are clearly discernible, consequently enhancing the differentiation between classes. Due to the use of layer norm in the CVV module, notable scale differences are observed in the statistical values of the two sets. Comparatively, the Euclidean distance between positive and negative classes relatively increased after the utilization of CVV adaptation. The effectiveness of our CVV module, as evidenced by the distribution results of positive and negative classes in the two sets of statistics, can be attributed to our incorporation of contrastive learning techniques. While some overlap between positive and negative samples persists even after CVV adaptation, it is reduced notably. Meanwhile, our geospatial relation discrimination relies on a neural network and doesn't solely depend on the Euclidean distance between samples.

5.2 Ablation of sharing weights in CVV

Although the backbone of our cross-view retrieval model shares weights across its two branches, in the feature adaptation process, the CVV modules in the two branches do not share weights. We hypothesize that the substantial difference in viewing angles leads to distinct feature representation between cross-viewing images. To validate the rationale behind our design, we established a control group where we configured the CVV modules of the two branches in the model to share weights, while keeping other components unchanged. We subsequently retrained the model and conducted tests on sub-datasets 11-15 to calculate the average Recall@1 and AP. The results are shown below.

Method	Recall@1	AP
ViT-S/14+1CVV	40.31	43.37
ViT-S/14+2CVV(ours)	64.50	68.67

Table 4. Comparison between sharing CVV weights and not sharing CVV weights

On the Recall@1 metric, the group that didn't share CVV weights demonstrated a 24.19% improvement compared to the control group that shared CVV weights, while AP increased by 25.30%. This disparity is quite evident, hence we opt not to share CVV weights in our method. Although our experimental results do not elucidate the relationship between feature representation and viewing angle, our design significantly enhances performance, as evidenced by the test outcomes.

5.3 Analysis of overfitting in the proposed model

Despite the promising performance on the sub-dataset, we acknowledge that our method still exhibits shortcomings when evaluated across the entire test set. During the training process, we observed that after a certain number of epochs, the recall on the validation and test sets ceased to improve, while the recall on the training set had already converged to a very high level. This suggests that our model may be experiencing overfitting. One of the most common and effective strategies is to mitigate overfitting by reducing the complexity of the network and opting for a more lightweight architecture. Therefore, we devised a control experiment wherein we eliminated the transformer block from the CVV module, retaining only the MLP head. This reduction in complexity significantly decreased the number of trainable parameters in the model from 2.8 million to 0.3 million. Subsequently, we retrained the modified model and evaluated its performance on sub-datasets 11-15, calculating the average Recall@1 and AP. The results are shown in the Tab. 5.

Method	Recall@1	AP
ViT-S/14+2MLP	49.37	54.81
ViT-S/14+2CVV(ours)	64.50	68.67

Table 5. Comparison between complete CVV and CVV removing transformer block (retaining only MLP)

Despite the considerable reduction in model complexity, there was a decline of over 10% in both Recall@1 and AP performance. During the training process of the control group model, we observed that it could still converge to very high recall on the training set. However, the recall plateaued at a low level on both the validation set and test set, indicating that for our model, there isn't a straightforward correlation between overfitting and model complexity. In addition, we have attempted various methods to address overfitting, including incorporating dropout, utilizing regularization techniques like Batch Normalization and L2 regularization, applying data augmentation, reducing the training batch size, etc. However, these measures did not significantly enhance performance on the test set. We speculate that this might be indicative of a non-standard form of overfitting, or potentially overfitting specifically to the training set.

5.4 Limitations of the proposed model

Our experiments validate that the visual foundation model is helpful for cross-view retrieval tasks. However, the representation performance of global features on the test set does not match that of the training set, indicating potential for further improvement in model generalization. In comparison with some of the latest cross-view image retrieval methods, although supervised learning is integrated, our supervision relies on positive and negative classes while discarding absolute position labels during the process. This approach constitutes relatively weak supervision, which may limit the model's ability to learn optimal features. Additionally, researchers have suggested that in contrastive learning, the generalization of contrastive loss might be slightly weaker than loss functions such as triplet loss. Hence, for future research, we propose improvements in three main areas: transfer learning, contrastive learning, and supervised learning. Specifically, we aim to enhance the CVV module to extract more generalizable global features, refine the loss function to further optimize the distribution of positive and negative class samples in feature space, and integrate absolute position information of cross-view images into supervised learning to reinforce geospatial supervision.

6. Conclusions

The retrieval of satellite images based on oblique UAV images poses a significant challenge due to the considerable difference in angles of views between the two types of imagery. To address this challenge, we propose a deep learning method based on a visual foundation model and ViT. Our method integrates transfer learning, contrastive learning, and supervised learning to tackle cross-view image retrieval tasks.

Most cross-view image retrieval methods based on deep learning often utilize a large and complex architecture, leading to considerable costs in training time and computation. This limitation severely restricts the practical application of these methods in engineering tasks. In our work, we explore the potential of the visual foundation model in cross-view image retrieval tasks. By leveraging prior knowledge acquired by DinoV2 on large-scale datasets, we achieve effective network initialization, and maintain the zero-shot transfer feature of DinoV2. Consequently, we only need to train the lightweight feature adaptation module and classification head, significantly reducing the complexity of the cross-view image retrieval model and enhancing the method's feasibility in engineering tasks.

Additionally, we harness ViT's capabilities in capturing global features, rendering our features more suitable for retrieval tasks.

Experiments on public datasets demonstrate that our method excels when the number of scene categories is under 50 and satisfactorily meets the requirements of cross-view image retrieval in engineering applications with streamlined scene categories and constrained computational resources. Nevertheless, there is still potential for improvement in the generalization ability of our method to non-training data. For future enhancements, we plan to focus on three aspects: feature adaption, loss function of contrastive learning, and supervised techniques.

Acknowledgements

This project was supported by the Natural Science Foundation of Shanghai (Grant No. 22ZR1465700)

References

Verykokou, S., Doulamis, A., Athanasiou, G., Ioannidis, C., Amditis, A., 2016. UAV-based 3D modelling of disaster scenes for Urban Search and Rescue. *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*. pp. 106-111, IEEE, 2016.

Sheppard, C., Rahnemoonfar, M., 2017. Real-time scene understanding for UAV imagery based on deep convolutional neural networks. *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. pp. 2243-2246. IEEE, 2017.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P., 2023. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193 (2023).*

Cheng, L., Yuan, Y., Xia, N., Chen, S., Chen, Y., Yang, K., ... & Li, M., 2018. Crowd-sourced pictures geo-localization method based on street view images and 3D reconstruction. *ISPRS journal of photogrammetry and remote sensing, 141*, pp. 72–85, 2018.

Luo, J., Ye, Q., 2023. UAV large oblique image geo-localization using satellite images in the dense buildings area. *Proceedings of 5th ISPRS Geospatial Week (GSW), Cairo, Egypt*, pp. 1065-1072.

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

Zemene, E., Tesfaye, Y. T., Idrees, H., Prati, A., Pelillo, M., Shah, M., 2018. Large-scale image geo-localization using dominant sets. *IEEE transactions on pattern analysis and machine intelligence, 41(1)*, pp. 148–161, 2018.

Rodrigues, R., Tani, M. 2022. Global assists local: Effective aerial representations for field of view constrained image geo-localization. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2694–2702, 2022.

Zhang, X., Meng, X., Yin, H., Wang, Y., Yue, Y., Xing, Y., Zhang, Y., 2021. SSA-Net: Spatial scale attention network for image-based geo-localization. *IEEE Geoscience and Remote Sensing Letters, 19*, pp. 1-5, 2022.

Lin, J., Zheng, Z., Zhong, Z., Luo, Z., Li, S., Yang, Y., Sebe, N., 2022. Joint representation learning and keypoint detection for cross-view geo-localization. *IEEE Transactions on Image Processing, 31*, pp. 3780-3792, 2022.

Zeng, Z., Wang, Z., Yang, F.,Satoh, S. I., 2022. Geo-localization via ground-to-satellite cross-view image retrieval. *IEEE Transactions on Multimedia, 25*, pp. 2176-2188, 2022.

Hu, S., Feng, M., Nguyen, R. M., Lee, G. H., 2018. CVM-Net: Cross-view matching network for image-based ground-to-aerial geo-localization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7258-7267, 2018.

Dai, M., Hu, J., Zhuang, J., Zheng, E., 2021. A transformer-based feature segmentation and region alignment method for uav-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology, 32(7)*, pp: 4376-4389, 2021.

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

Zhou, C., Li, Q., Li, C., Yu, J., Liu, Y., Wang, G., ... & Sun, L., 2023. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. *arXiv preprint arXiv:2302.09419 (2023).*

Doersch, C., Gupta, A., Efros, A. A., 2015. Unsupervised visual representation learning by context prediction. *Proceedings of the IEEE international conference on computer vision*, pp. 1422-1430, 2015.

Bao, H., Dong, L., Piao, S., Wei, F., 2021. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254 (2021).*

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... & Girshick, R., 2023). Segment anything. *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015-4026, 2023.

Ibrahimovic, E., 2023. Optimizing Vision Transformer Performance with Customizable Parameters. *2023 46th MIPRO ICT and Electronics Convention (MIPRO), Opatija, Croatia*, pp. 1721-1726, IEEE, 2023.

Lu, F., Lan, X., Zhang, L., Jiang, D., Wang, Y., Yuan, C., 2024. CricaVPR: Cross-image Correlation-aware Representation Learning for Visual Place Recognition. *arXiv preprint arXiv:2402.19231 (2024).*

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., ... & Gelly, S., 2019. Parameter-efficient transfer learning for NLP. *In International Conference on Machine Learning*, pp. 2790–2799. PMLR, 2019.

Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290-1299. 2022.

Zhou, W., Guan, H., Li, Z., Shao, Z., Delavar, M. R., 2023. Remote Sensing Image Retrieval in the Past Decade: Achievements, Challenges, and Future Directions. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16*, pp.1447-1473, 2023.