SUGISAKA MASANORI

Journal of Robotics, Networking and Artificial Life
Vol. 9(3); December (2022), pp. 275–284
ISSN (Online):2352-6386; ISSN (Print):2405-9021
https://alife-robotics.org/jrnal.html

Research Article

A Part-aware Attention Neural Network for Cross-view Geo-localization between UAV and Satellite

Duc Viet Bui, Masao Kubo, Hiroshi Sato
Department of Computer Science, National Defense Academy, 1-10-20 Hashirimizu, Yokosuka City, Kanagawa Prefecture, Japan

ARTICLE INFO

Article History
Received 29 June 2022
Accepted 20 September 2022

Keywords
Cross-view image matching
Geo-localization
UAV
Attention mechanism
Part-based representation learning

ABSTRACT

Cross-view image matching for geo-localization is the task of finding images containing the same geographic target across different platforms. This task has drawn significant attention among researchers due to its vast applications in UAV's self-localization and navigation. Given a query image from UAV-view, a matching model can find the same geo-referenced satellite image from the database, which can be used later to precisely locate the UAV's current position. Many studies have achieved high accuracy on existing datasets, but they can be further improved by combining different feature processing methods. Inspired by previous studies, in this paper, we proposed a new framework by using a channel-based attention mechanism combined with a part-based representation learning method, including multi-level feature aggregation and an alternative pooling strategy to enhance the feature extracting process. The proposed model significantly improved matching accuracy and surpassed the existing state-of-the-art methods on University-1652 dataset..

1. Introduction

The applications of unmanned aerial vehicles (UAV) in daily life have been rapidly increasing. The UAV has become an essential part of various fields such as aerial surveillance [1], agriculture [2], transportation, and search and rescue missions. Along with their applications, to further reduce human's work, the need for autonomous drones has been increasing over time. However, most researches failed to achieve a fully autonomous drone system, as the most used navigation system (Global Positioning System - GPS) has many limitations in real-life missions. For example, GPS is not powerful enough when high buildings or trees block GPS signals, leading to difficulties in applying UAVs in cities and urban areas. Many solutions for navigation in autonomous drone systems have been proposed, and among them, cross-view image matching-based methods have received lots of researchers' attention due to the vast application value in geo-localization [3-6]. Cross-view image matching is the task of matching a satellite-view image with geographic location tags and a UAV-view image without a geographic location tag or vice versa to locate a UAV's position based on information from taken images. Figure 1 shows an example of cross-view matching methods. Given a UAV-view image of a building, the matching model searches for the image of that building in the satellite-view image gallery. The output is a satellite-view image similar to the query UAV-view image. This output can be used to locate the current position of the UAV.

Corresponding author's E-mail: ed21009@nda.ac.jp, masaok@nda.ac.jp, hsato@nda.ac.jp

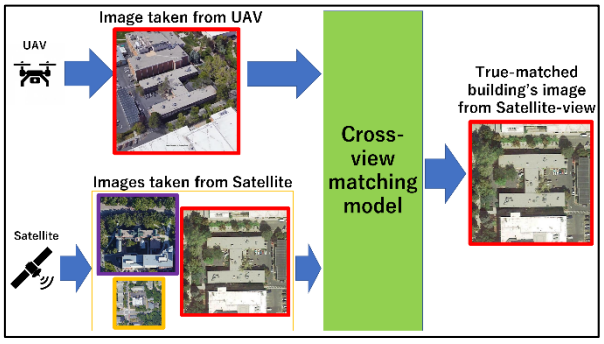

Fig. 1: Example of UAV-view → Satellite-view

Early-stage cross-view image matching researches [7] focused on using traditional image processing methods, which used hand-crafted features from images. In recent years, with the rapid development of machine learning and deep learning in image processing, especially the convolutional neural network (CNN), many studies attempted to apply them in cross-view image matching for geo-localization, some of which have achieved significant results [3,6]. Also, deep learning's well-known self-attention mechanisms have been used in several cross-view geo-localization studies [8-10] to understand image representations further and bridge the gap between images from different views. However, the self-attention mechanisms are rather complicated and usually require lots of computational costs, which may not be an ideal method for current UAV systems. Another branch of research focuses on learning representations based on feature parts; an approach derived from Person Re-Identification (Person Re-ID) related research. These methods divide feature maps into small part and help the models learn sub-salience features in the images. However, the feature partition phase in previous works usually process only the final feature map of the model and ignore the features from shallow layers. Geographic targets in the dataset are mainly buildings and roads, so we consider low-level features extracted from shallow layers may also play an important role in understanding the entire view.

Therefore, different from previous cross-view geo-localization works which rarely applied attention mechanisms and ignored the importance of part-based representation learning towards low-level features, in this paper we proposed a new framework which used a channel-based attention mechanism and implemented a part-based representation learning with multi-level feature aggregation, and also an alternative pooling strategy. Our proposed model has shown an increase in performance through experiments compared to the state-of-the-art method (SOTA) and other existing methods.

2. Related works

2.1. Cross-view Image Matching for Geo-localization

Previous studies [11][12] in cross-view image matching often consider this task as an image retrieval problem since they aim to find similar images to a query image among an image dataset. In the cross-view matching problem, the main task is to learn image representation that varies in different views, thus bridging the gap between multi-view images. The schematic diagram of cross-view matching is described in Figure 2. At first, features from query images and a database (gallery images) are extracted using different feature processing methods. After that, features' similarities were calculated by using distance similarity measures such as cosine similarity or Euclidean distance. The results are later used to create a ranking list, from which the model will determine the true-matched image to the query image.

The methods used in cross-view geo-localization related works can be categorized as below.

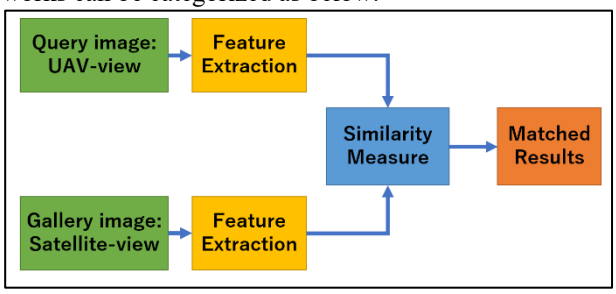

Fig. 2: The schematic diagram of cross-view image matching problem (UAV-view → Satellite-view)

Feature Extracting Methods Traditional image processing methods such as SIFT [13], or SUFT [14] have been implemented in early research for the feature extracting phase. However, as the gap between different viewpoints is enormous, the images of the same location from different views are dramatically different. As a result, direct image matching using traditional methods tends to fail. Recently, more and more researchers have been paying attention to the powerful convolutional neural network (CNN), which is well-known for its ability to learn high-level features. Workman et al. [15] were the first ones that attempted to apply a pre-trained CNN (pre-trained AlexNet on ImageNet dataset) to this

problem. Their results proved that the features learned by this method outperformed hand-crated features from traditional methods. Furthermore, in [16], Workman et al. continued to increase the matching accuracy by reducing the feature distance between pairs of aerial images and ground-view images, which boosted the pre-trained model's performance. Hu et al. [17] adapted NetVLAD (which is a novel CNN model for image retrieval) into a cross-view matching model and achieved competitive results on the famous CVUSA [16] dataset. Zhai et al. [18] also modified NetVLAD and Siamese network architecture to capture the semantic layout of satellite-view images, which robust the image descriptor in retrieving images. Ding et al. [19] considered this problem as a place-classification problem and developed a ResNet50-based model to solve it.

Metric Learning Loss Another line of work focused on learning discriminative features using metric learning and proposing various loss functions. This line of work has similar approaches with face verification and Person Re-ID problems cause they adopt ranking losses such as contrastive loss or triplet loss to learn the relative distances between different inputs. By mapping relevant features onto their appropriated spaces and learning relative distances between different spaces, matching models can discriminate images that they have not seen before. Inspired by these approaches, Lin et al. [20] applied Siamese network architecture and adopted contrastive loss to optimize network parameters. Vo and Hays [21] proposed soft-margin ranking loss, which was an attempt to overcome the margin issues of margin triplet loss. Hu et al. [17] improved soft-margin ranking loss by introducing weighted soft-margin ranking loss, which further reduced convergence in a training phase. Different from other works which adopted common ranking loss, several researchers [22-24] applied the instance loss [25], which is inspired by classification loss, and achieved remarkable retrieval results.

Attention Mechanism Moreover, some works try to enhance the feature learning phase by applying the attention mechanism. Attention-mechanism is a method invented to make neural networks learn the most relevant features from inputs, thus increasing the network's learning abilities. Emphasized features extracted by attention-mechanism tend to contribute positively towards the final prediction result. For example, Shi et al. [26] used spatial attention mechanism to enhance the performance of cross-view geo-localization model. Recently, novel self-attention model in natural language processing – Transformer [27] has been applied in in numerous of vision processing researches (known as Vision Transformer [28]) as well as cross-view geo-localization related works [8,29,30], and received some positive results. However, the disadvantages of Transformer architecture are high computational cost and a massive amount of data required for training. Famous cross-view image datasets [9][16][22] often contain a little amount of images, thus it may be difficult to properly train a Transformer-based model on these datasets.

2.2. Part-based Representation Learning

Several studies in fields of computer vision have paid attention to part-based representation learning, a feature processing method that divides the feature maps into small feature parts and supervise them. The fine-grained information from splitted parts are expected to help models understand comprehensive features of the entire image. This method is often seen in Person Re-ID researches, as splitting features of human body parts and aligning them can help the model extract high-level segmentation features. In 2018, the Part-based Convolutional Baseline (PCB) [31] greatly surpassed the SOTA of Person Re-ID problem by horizontally divided and matched human body's feature maps. Following the idea of PCB, Luo et al. [32] proposed AlignedReID++ which also aligned sliced feature parts to jointly learn the global features and local features. In Cross-view geo-localization, this technique recently becomes popular: LPN [24] and MBSA [33] invented square-ring feature partition strategies, which encouraged the network to pay more attention to fine-grained information from the edge of input images. The idea were continuously developed in pixel-level: by using VisionTransformer as backbone, FSRA [29] and SGM [30] models divided feature maps into pixels and re-arranged them based on each pixel's attributes. These achievements in cross-view geo-localization show promising results in applying part-based representation learning in cross-view problems.

2.3. Multi-level Feature Aggregation

Features from various layers offer varying degrees of semantic information. Merging the feature maps from

multiple layers in CNN can result in enhanced feature discrimination ability. In general, shallow layer features often extract image local structures and fine-grained information such as shapes and edges, but they can not represent global semantic information and usually contain noises. Deeper layer features may provide high-level global meanings, but lack spatial and detailed information. As a result, this technique is commonly employed in numerous computer vision applications. Multi-level feature aggregation can be implemented by developing multiple branches on multi-scale and different parts of the model. For example, Kirillov et al. [34] fused element-wise operations with different scales of feature maps; Li et al. [35] combined multiple branches of a network to combine local and global features, in order to extract human feature representation.

3. Materials and Proposed Method

3.1. Dataset and Evaluation Metrics

In this work, we use the University-1652 dataset published by Zheng et al. [22], as it is the only benchmark dataset with both satellite-view and UAV-view images, which helps solving cross-view geo-localization for UAV navigation. This dataset contains 1652 geographic targets from 72 universities all over the world. Each target contains three views: satellite-view, UAV-view, and street-view. To reduce the high cost in airspace control and flying UAV, all UAV-view and street-view images were collected by a 3D engine named Google Earth, while satellite-view images were captured by Google Map. All images in the dataset have geo-tags as their class labels. The view of UAVs in Google Earth was controlled by simulated camera-view, and the height of view descends 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 and test sets with no overlapped classes. The captured images have an original size of 512 × 512. The distribution of data in each set is described in Table 1. Samples of images in the dataset were demonstrated in Figure 3.

Table 1. Distribution of image data in University-1652.

	Number of Images	Number of classes	Number of Universities
Training	50218	701	33
Query (UAV)	37855	701	39
Query (Satellite)	701	701	
Query (Ground)	2579	701	
Gallery (UAV)	51355	701	
Gallery (Satellite)	951	951	
Gallery (Ground)	2921	793	

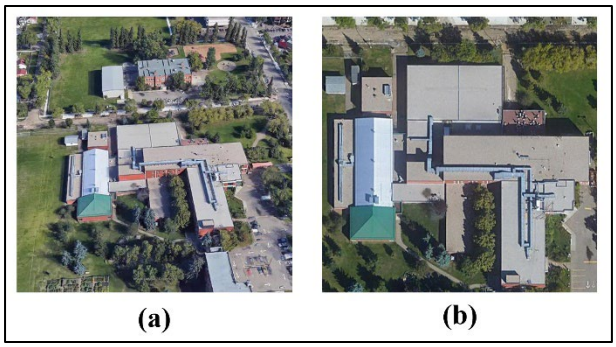

Fig. 3: Building's images from UAV-view (a) and Satellite-view (b) in University-1652.

Regarding the evaluation metrics, most of the image retrieval and cross-view image matching researches has been using Recall@K and Average Precision (AP) as the main indicator for evaluating proposed systems. 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. The higher Recall@K and AP, the better the model performs.

3.2. Proposed Method

The overview of proposed network architecture is described in Figure 4. The network was divided into two branches for each input; in each branch we deploy our proposed feature extractors, and they share the same intial weight. Extracted features from each branch were sent to a Classifier module, which is composed of multiple Fully Connected layers (FC) and Classifier layers (Cls). From subsection 3.2.1 to 3.2.3, we explain the details of our proposed feature extractors in Figure 5.

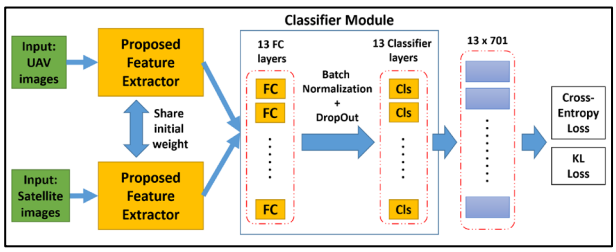

Fig. 4: Proposed network architecture

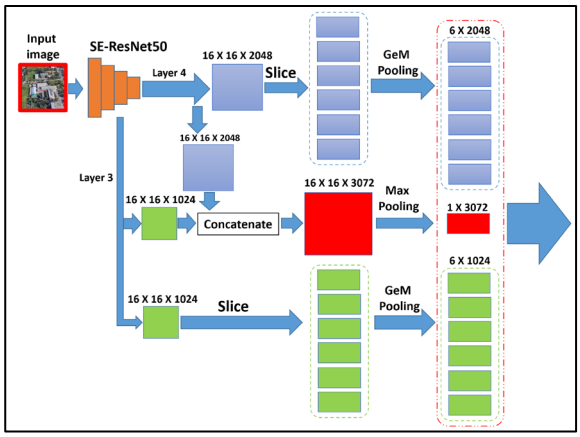

Fig. 5: Proposed feature extractor

3.2.1. *Attention-based Feature Extractor*

Because of excellent accuracy and inference time, other existing methods used CNN backbone from ResNet50 [36] model or VGG16 [37] as the main feature extractor, while the usage of attention modules in these backbones is rarely seen. However, we believe that an attention mechanism can strengthen the saliency value of each view and restrain the unnecessary features from affecting the final results. Therefore, among various attention mechanisms in literature, channel-based attention - the SE-block [38] was chosen for its efficiency in reinforcing the backbone while requiring very little additional computation cost. The SE-block performs channel reduction process, which helps re-adjusting the weight of each channel and emphasizing meaningful channels.Especially, the SE-block could be easily implemented in ResNet50's Residual blocks as described in Figure 6.

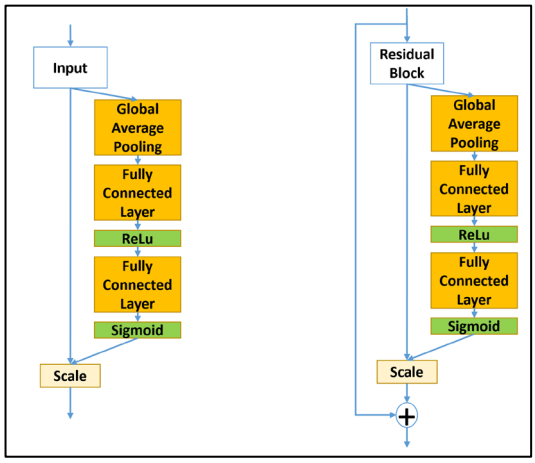

Fig. 6: SE block (left) and its implementation in Residual block (right)

3.2.2. *Part-aware Multi-level Feature Fusion*

Previous works that applied part-based representation learning in [24] and [33] proposed a brand new feature partition strategy to take advantage of contextual information. In particular, the output feature map is divided into several parts called square-ring blocks, and then the Global Average Pooling method is performed. Here we also applied the feature partition strategy that creates multi-scale square-ring blocks, which is described in Figure 7.

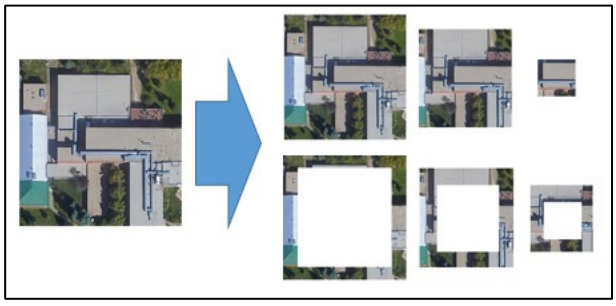

Fig. 7: Multi-scale block partition

Especially, previous works only applied this strategy to features from the fourth-layer group, while in our proposed model the features from other layers (here, we chose the third-layer group) were also extracted, and the square-ring feature partition strategy was applied to both types of features (the blue-color and green-color boxes in Figure 5). Furthermore, in previous studies [33], fusion of multi-level features contributed greatly to the final results; thus, here we also created a global feature map which was concatenated by the results of third and fourth-layer groups (the red-color box in Figure 5). Notice that the stride of the final down-sampling layer

was fine-tuned from 2 to 1 so the features from the fourth-layer group can have the same size with features from the third-layer group.

3.2.3. *Pooling strategy*

Different from related studies which applied Global Average Pooling, in this work Generalized Mean Pooling (GeM pooling) was put in practice. GeM pooling was first proposed in [39] as an alternative pooling method for image retrieval. Since then, it has been widely applied in many retrieval systems and achieved promising results. The formula of GeM pooling can be defined as follow:

$$f^{(g)} = [f_1^{(g)} .. f_k^{(g)} .. f_K^{(g)}]^T, f_k^{(g)} = (\frac{1}{|X_k|} \sum_{x \epsilon X_k} x^{p_k})^{\frac{1}{p_k}} \quad (1)$$

with X_k represents feature map, K is the number of channel and p_k is the pooling parameter. Notice that this pooling parameter is a learnable parameter: it can be manually set or changed through learning process. For the concatenated feature map of third and fourth-layer group, we applied the Global Max Pooling.

3.2.4. *Loss Functions and Learning Strategy*

In [22-24], instance loss was adopted to train the multi-branches networks, and results have shown the good effect of this loss function compared to other ranking losses in terms of cross-view matching accuracy. Instance loss was first proposed in [25], an alternative way to learn the distance between features. Extracted features from each branch were sent to the shared fully connected layer in order to map the features of all sources into one shared feature space. Finally, the Cross-Entropy loss function was applied to optimize the network.

Additionally, in [33], Kullback-Leibler divergence (KL divergence) was first applied in the training phase. In the fields of machine learning, KL divergence is a measure of how a probability distribution differs from another probability distribution. In this problem, KL divergence is expected to close the gap between two different domains (UAV and Satellite).

Here we used the Softmax function to obtain the normalized probability scores, and KL divergence was then computed and added to the training loss. The KL divergence loss formula is defined as follow:

$$L_{KL(p_2||p_1)} = \sum_{n=1}^{N} p_2^n log \frac{p_2^n}{p_1^n} \quad (2)$$

p_1, p_2 are predicted results of each branch.

3.2.5. *Implementation details*

For training, we resize all the input images to the size of 256 × 256. Random flipping and random cropping were used to augment the input images before training. SE-ResNet50 was pre-trained with the ImageNet dataset. The optimizer was Stochastic Gradient Descend (SGD) with a momentum of 0.9 and weight decay of $5x10^{-4}$. The training lasted for 120 epochs, with an initial learning rate of $1x10^{-4}$ for backbone layers and $1x10^{-3}$ for other layers. The pooling parameter p_k in GeM pooling was initially set to 3. In the testing phase, the classifier layers was removed so the model returns only extracted features as outputs. Euclidean distance was applied to compute the similarity between feature vectors from different views. We used Recall@1 (R@1) and AP for evaluating models' performance. To investigate the calculation requirement of the proposed method compared with others, we calculated the parameters that each model contained and the inference time required for retrieving one picture at a time.

4. Experiments and Discussions

We performed training our proposed model with the University-1652 dataset and some ablation experiments to understand the effectiveness of our model and learning strategies towards the results.

4.1. Comparison with SOTA and other methods

In Table 2, we compared our proposed method with the SOTA in [33] and other related works. Our model achieved 84.51% R@1 accuracy and 86.78% AP on Drone → Satellite, 91.01% R@1 accuracy and 82.28% AP on Satellite → Drone data. The performance of the method greatly surpassed all the existing competitive

Table 2. Comparison of proposed method with SOTA

Method	Backbone	University-1652				Parameters	Inference Speed
		UAV→Satellite		Satellite→UAV			
		R@1	AP	R@1	AP		
Baseline [22]	ResNet-50	58.23	62.91	74.47	59.45	26M	1.00 ×
SFPN [23]	ResNet-50	58.49	63.13	71.18	58.74	26M	1.00 ×
LCM [19]	ResNet-50	70.29	73.88	79.74	69.40	26M	1.00 ×
LPN [24]	ResNet-50	75.93	79.14	86.45	74.49	26M	1.00 ×
LPN [24]	ResNet-101	76.13	79.29	85.45	75.45	45M	1.51 ×
FSRA [29]	Vision Transformer	84.51	86.71	88.45	83.37	51M	1.05 ×
SGM [30]	SwinTransformer	82.14	84.72	88.16	81.81	28M	1.04 ×
MBSA [33]	ResNet-50	82.33	84.78	90.58	82.06	36M	1.05 ×
Ours	**SE-ResNet50**	**84.51**	**86.78**	**91.01**	**82.28**	**43.8M**	**1.07 ×**

models, including the SOTA (MBSA network). Especially, this model with channel-based attention backbone exceed other works that used novel Transformer-based backbones. For the calculation cost, although our proposed model was complicated and contained a huge amount of parameters (about 43.8 million), the inference time did not increase much compared to other methods (about **1.07** × compared to the baseline method).

4.2. Ablation studies

4.2.1. *SE-block in comparison with other attention mechanisms*

To understand the effectiveness of attention mechanism used in this work (SE-block) towards the final results, we performed several experiments with different attention modules which shared the same concept with SE-block. BAM (Bottleneck Attention Module) [40] and CBAM (Convolutional Block Attention Module) [41] were chosen as the comparison targets. Notice that all these modules performed better than SE-block in the Large Scale Visual Recognition Challenge (ILSVRC) [42]. The proposed framework was trained with different attention-based backbones and the same training conditions. As shown in Table 3, for Satellite → UAV (UAV → Satellite), the R@1 and AP of the SE-ResNet50-based method are higher than ResNet50-based by 1.86% (2.42%) and 1.77% (1.13%), respectively. Additionally, SE-block achieved the best performance compared to ResNet50 other attention-based methods. We assume that the attention weight screated by the channel reduction process of SE-block has emphsized important parts of entire feature maps and greatly made impact on the final results.

Table 3. Comparison of different attention-based backbone

Backbone	University-1652			
	UAV → Satellite		Satellite → UAV	
	R@1	AP	R@1	AP
ResNet50	82.65	85.01	88.59	81.15
BAM-ResNet50 [40]	79.50	82.27	87.73	78.06
CBAM-ResNet50 [41]	83.95	86.32	90.44	82.23
SE-ResNet50	**84.51**	**86.78**	**91.01**	**82.28**

4.2.2. *Comparison of different part-aware multi-level feature*

Does all the feature from shallow layers important? Should we applied the feature partition strategy to all features from different layers? To answer these question, ablation experiments with different levels of feature were executed. We use the number 1, 2, 3 and 4 to represent which layers group of SE-ResNet50 was applied feature partition strategy. For example, (3 + 4) in Table 4 means that the features from third and fourth-layer group of SE-ResNet50 were applied partition strategy (which is also our proposed model in Section 3). Table 4 demonstrated the experiment results; R@1 and AP drop dramatically when using features from first and second-layer group, while the combination of third and fourth-layer group achieved the best performance. We assume that features from first and second-layer group are still not capable of represent the meaning of the entire image; thus using them in the final feature map could make the model behave badly.

Table 4. Comparison of different part-aware multi-level feature

Partition Feature on Layer	University-1652			
	UAV → Satellite		Satellite → UAV	
	R@1	AP	R@1	AP
4	82.87	85.13	90.87	82.06
3 + 4	**84.51**	**86.78**	**91.01**	**82.28**
2 + 3 + 4	82.49	84.93	88.30	78.95
1 + 2 + 3 + 4	77.70	80.67	85.02	75.32

4.2.3. *Comparison of different pooling strategies*

In CNN, pooling method is the key to extract meaningful feature and discard irrelevant information. To investigate the influence of pooling strategies in our proposed method, several experiments with different combinations of pooling methods were conducted. In Table 5, **Local** column demonstrated the pooling method used for partition features from third and fourth-layer group, while **Global** column described the pooling method used for the concatenated feature.

The experiments' results in Table 5 demonstrated that among all the pooling methods, features extracted by GeM Pooling outperformed other pooling methods in both tasks. This confirmed the power of GeM Pooling in image retrieval tasks. However, when applied GeM Pooling to the concatenated global feature and local partition features at the same time, the accuracy dropped off nearly 1~2%. From these results, we assume that GeM Pooling performed well in enhancing features created by part-based representation learning, while Max Pooling may be the better for generalizing concatenated features. These findings could be used as references for future works which involving part-aware multi-scale features.

Table 5. Comparison of proposed method with SOTA

Pooling Strategy		**University-1652**			
		UAV → Satellite		**Satellite → UAV**	
Local	**Global**	**R@1**	**AP**	**R@1**	**AP**
Avg	Avg	83.29	85.75	88.87	80.41
	Max	82.78	85.33	89.30	80.79
	GeM	82.72	85.19	90.01	80.94
Max	Avg	49.68	54.20	65.34	50.52
	Max	53.26	57.79	67.33	53.35
	GeM	56.43	60.86	69.04	54.29
GeM	Avg	84.18	86.49	89.44	81.19
	Max	**84.51**	**86.78**	**91.01**	**82.28**
	GeM	83.81	86.18	89.59	82.26

4.2.5. *Comparison of different input image sizes*

In general, low-resolution images may negatively affect the model's performance but increase the inference time, while high-resolution images give models better information but take more time for models to process. The trade-off between accuracy and speed is a big challenge in developing real-world applications with limited computing resources, especially in UAV-related applications. Here we used some ablation experiments to observe the changes of proposed model in terms of input image size. From the results in Table 6, when the resolution increases, proposed model's performance also grows. However, in Satellite → Drone task, when the resolution changed from 384 to 512, the performance decrease a little. These results show that the proposed model robursts to the change of resolution, which is useful for future UAV applications when selecting input image size.

Table 6. Comparison of different input image sizes

Method	**University-1652**			
	UAV → Satellite		**Satellite → UAV**	
	R@1	**AP**	**R@1**	**AP**
224 × 224	78.75	81.57	85.45	76.91
256 × 256	84.51	86.78	91.01	82.28
384 × 384	87.05	88.96	**92.44**	**84.16**
512 × 512	**87.79**	**89.56**	90.87	85.14

4.3. Qualitative results on 4K-images

The University-1652 dataset also provided a small amount of 4K-resolution images, which were collected by real drones flying above several university mentioned in University-1652 dataset. Compared to the training-testing data that were simply collected in 3D simulation, these images can be considered as real-world images. To confirm the reliability of our proposed framework for real-world mission, we visualized the results created by our model on these 4K-resolution images in Figure 8 (Real UAV-view image→ Satellite-view image). The true-matched images are in yellow boxes, and the false-matched images are in blue boxes. Although the Satellite gallery only contains one image for each place, which is rather hard to find the correct answer, the proposed model can still find the accurate matching image. This result proved that even our proposed model was trained on simulation images, it could perform greatly on real images, which shows the capability of applying this model in a real UAV missions.

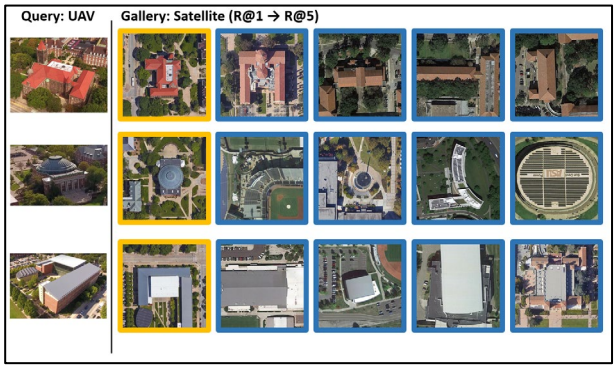

Fig. 8: Top-5 retrieval results of Real UAV-view image → Satellite-view image

5. Conclusion

Recent years, with the advance of UAV technology, the need for autonomous control of UAV is increasing rapidly, especially navigating UAV withou GPS signals. In this paper, we addressed this problems as a vision

processing task in UAV and focused on solving cross-view image matching tasks for geo-localization. We revealed the shortcoming of existing methods, and designed a new architecture using a channel-based attention network as feature extractors and a part-aware multi-level feature learning strategy. The performance of model was verified on a benchmark dataset (University-1652). Experiment results showed that our proposed model has an increase in accuracy compared to the previous SOTA and other existing methods. As shown in ablation experiments, each component in our model contributed positively towards the final matching results. In future works, to increase the model's robustness to features in cross-view domains, we plan to further exploit the partition features from shallow layers of the network. Utilizing the network to match the requirements of limited computing resource will also be a challenge in the next phase of research.

References

1. M. Kontitsis, K. P. Valavanis and N. Tsourveloudis: "A UAV vision system for airborne surveillance", IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004, New Orleans, LA, USA, pp. 77-83 Vol.1, 2004.
2. S. W. Chen, et al: "Counting apples and oranges with deep learning: a data-driven approach", IEEE Robotics and Automation Letters, vol. 2, no. 2, 2017.
3. L. Ding, et al.: "A Practical Cross-View Image Matching Method between UAV and Satellite for UAV-Based Geo-Localization.", Remote Sensing 13.1: 47, 2021.
4. L. Liu, H. Li, and Y. Dai.: "Stochastic Attraction-Repulsion Embedding for Large Scale Image Localization", Proceedings of the IEEE International Conference on Computer Vision, pp. 2570–2579, 2019.
5. Y. Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. : "Optimal Feature Transport for Cross-View Image Geo-Localization". arXiv preprint arXiv:1907.05021 , 2019.
6. Y. Tian, C. Chen, and Mubarak Shah: "Cross-view image matching for geo- localization in urban environments", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3608–3616, 2017.
7. F. Castaldo et al.: "Semantic cross-view matching", Proceedings of the IEEE International Conference on Computer Vision Workshops, p. 9-17, 2015.
8. H. Yang et al.: "Cross-view Geo-localization with Evolving Transformer",arXiv preprint arXiv:2107.00842, 2021.
9. L. Liu and Hongdong Li: "Lending orientation to neural networks for cross-view geo-localization", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
10. L. Liu and Hongdong Li: "Lending orientation to neural networks for cross-view geo-localization", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019.
11. N. Khurshid et al.: "Cross-View Image Retrieval - Ground to Aerial Image Retrieval Through Deep Learning", Proceeding of International Conference on Neural Information Processing, Springer, p. 210-221, 2019.
12. S. Zhu et al.: "VIGOR: Cross-View Image Geo-localization beyond One-to-one Retrieval", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, p. 3640-3649, 2021.
13. D. G. Lowe et al.: "Object recognition from local scale-invariant features", Proceedings of the IEEE International Conference on Computer Vision, 1999.
14. H. Bay, T. Tuytelaars, and L. Van Gool.: "Surf: Speeded-up robust features", Proceedings of European Conference on Computer Vision, 2006.
15. S. Workman and N. Jacobs: "On the location dependence of convolutional neural network features," Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015.
16. S. Workman, Richard Souvenir and Nathan Jacobs: "Wide-area image geolocalization with aerial reference imagery", Proceedings of the IEEE International Conference on Computer Vision, 2015.
17. S. Hu et al.: "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.
18. M. Zhai, Z. Bessinger, S. Workman, and N. Jacobs: "Predicting ground-level scene layout from aerial imagery", Proceeding of IEEE Conference on Computer Vision and Pattern Recognition, July 2017.
19. L. Ding, L., Zhou, J., Meng, L., & Long, Z.: "A practical cross-view image matching method between UAV and satellite for UAV-based geo-localization", Remote Sensing, 13(1), 2020.
20. T. Lin et al.: "Learning deep representations for ground-to-aerial geolocalization.", Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.
21. N. N. Vo and J. Hays: "Localizing and orienting street views using overhead imagery", Proceedings of European Conference on Computer Vision, 2016.
22. Z. Zheng, Y. Wei, and Y. Yang. : "University-1652: A multi-view multi-source benchmark for UAV-based geo-localization.", Proceedings of the 28th ACM international conference on Multimedia, 2020.
23. H. Sijin and Y. Wang: "Cross-view geo-localization via Salient Feature Partition Network", Journal of Physics: Conference Series, Vol. 1914, 2021.
24. W. Tingyu, et al. : "Each part matters: Local patterns facilitate cross-view geo-localization.", IEEE Transactions on Circuits and Systems for Video Technology, 2021.
25. Z. Zheng, et al.: " Learning Image-Text Embeddings with

Instance Loss" arXiv preprint arXiv:1711.05535, 2017.
26. Y. Shi, X. Yu, L. Liu, T. Zhang, and H. Li: “Optimal feature transport for cross-view image geo-localization,” Proceedings of AAAI Conference on Artificial Intelligence, 2020.
27. V. Ashish, et al.: “Attention is all you need”, Advances in neural information processing systems. pp. 5998-6008, 2017.
28. A. Kolesnikov et al. “An image is worth 16x16 words: Transformers for image recognition at scale”, arXiv preprintarXiv:2010.11929,2020.
29. M. Dai, J. Hu, J. Zhuang and E. Zheng: "A Transformer-Based Feature Segmentation and Region Alignment Method For UAV-View Geo-Localization," IEEE Transactions on Circuits and Systems for Video Technology, 2022.
30. 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.
31. Y. Sun, L. Zheng, Y. Yang, Q. Tian, and S. Wang: “Beyond part models: Person retrieval with refined part pooling (and a strong convolutional baseline),” Proceedings of the European conference on computer vision (ECCV), pp. 480–496, 2018.
32. H. Luo, W. Jiang, X. Zhang, X. Fan, J. Qian, and C. Zhang: “Alignedreid++: Dynamically matching local information for person reidentification,” Pattern Recognition, vol. 94, pp. 53–61, 2019.
33. J. Zhuang et al.: “A Faster and More Effective Cross-View Matching Method of UAV and Satellite Images for UAV Geolocalization”, Remote Sensing, 13.19: 3979, 2021.
34. A. Kirillov, R. Girshick, K. He and R. Dollár: “Panoptic feature pyramid networks,” Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6399-6408, 2019.
35. W. Li, X. Zhu and S. Gong: “Harmonious attention network for person re-identification”, Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2285-2294, 2018.
36. H. Kaiming, et al. : "Deep residual learning for image recognition.", Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.
37. K. Simonyan, Z. Andrew. : “Very Deep Convolutional Networks for Large-Scale Image Recognition”, arXiv 1409.1556, 2014.
38. J. Hu, L. Shen, and G. Sun.: "Squeeze-and-excitation networks.", Proceedings of the IEEE conference on computer vision and pattern recognition, 2018.
39. F. Radenović et al.: “Fine-tuning CNN image retrieval with no human annotation”, IEEE transactions on pattern analysis and machine intelligence, 41(7), pp 1655-1668, 2018.
40. J. Park et al.: "Bam: Bottleneck attention module." arXiv preprint arXiv:1807.06514, 2018.
41. S. Woo et al.: "Cbam: Convolutional block attention module." Proceedings of the European conference on computer vision (ECCV), 2018.
42. J. Deng: “Large scale visual recognition”, Diss. Princeton University, 2012.

Authors Introduction

Mr. Duc Viet Bui

He received his M.S degrees from Department of Computer Science, National Defense Academy of Japan in 2021. He is currently a doctoral student at Department of Computer Science in National Defense Academy of Japan. His research related to different applications of computer vision in aerial robotics. His interests include computer vision, machine learning, deep neural networks and aerial robotics.

Dr. Masao Kubo

He is Associate Professor of Department of Computer Science at National Defense Academy in Japan. He graduated from the precision engineering department, Hokkaido University, in 1991. He received his Ph.D. degree in Computer Science from the Hokkaido University in 1996 (multi-agent system). He had been the research assistant of the chaotic engineering Lab, Hokkaido university. He was a visiting research fellow of Intelligent Autonomous Lab, university of the west of England (2005). He is the associate professor of information system lab, Dep. of Computer Science, National Defense Academy, Japan. His research interest is Multi agent system.

Dr. Hiroshi Sato

He is an Associate Professor of the Department of Computer Science at the National Defense Academy in Japan. He obtained a degree in Physics from Keio University in Japan and degrees of Master and Doctor of Engineering from Tokyo Institute of Technology in Japan. He was previously a Research Associate at the Department of Mathematics and Information Sciences at Osaka Prefecture University in Japan. His research interests include agent-based simulation, evolutionary computation, and artificial intelligence. Dr. Sato is a member of the Japanese Society for Artificial Intelligence (JSAI), Society of Instrument and Control Engineers (SICE), and The Institute of Electronics, Information and Communication Engineers. IEICE). He was the editor of IEICE and SICE.