Learning Robust Feature Representation for Cross-View Image Geo-Localization

Wenjian Gan, Yang Zhou, Xiaofei Hu, Luying Zhao, Gaoshuang Huang, and Mingbo Hou

***Abstract*— The cross-view image geo-localization (CVGL) refers to determining the geographic location of a given query image using an image database with the known location information. Existing methods mainly focus on learning discriminative image representations to optimize the distance of image feature representations in feature space without fully considering the positional relation information of the features and the information redundancy in the features themselves. Therefore, we proposed a cross-view image localization method that combines the global spatial relation attention (GSRA) with feature aggregation. First, we utilize the lightweight GSRA to learn the spatial location structure information of features, which fully enhances the perceptual and discriminative capabilities of the model. The proposed attention has a little effect on the complexity and memory occupancy of the model and can be generalized to other image-processing tasks. In addition, we introduce the sinkhorn algorithm for locally aggregated descriptors (SALADs), which represents the aggregation of local features as an optimal transport problem and selectively discards useless information during the clustering and assignment of features, thus enhancing the generalization and robustness of the descriptors. Experimental results on the public University-1652, CVACT, and CVUSA datasets validate the effectiveness and superiority of the proposed method. Our code is available at: https://github.com/WenjianGan/LRFR.**

***Index Terms*— Cross-view image geo-localization (CVGL), feature representation, global spatial relation attention (GSRA), sinkhorn algorithm for locally aggregated descriptors (SALADs).**

I. Introduction

CROSS-VIEW image geo-localization (CVGL) refers to the technique of determining the geographic location in a preconstructed satellite reference image database for a given street-view or drone image [1], which is usually regarded as an instance image retrieval and has essential applications in automated driving [2] and drone localization [3]. However, the viewpoint variation and appearance differences between the query and database images make the research and application of CVGL a considerable challenge.

Early methods [4], [5] used hand-crafted features to achieve CVGL, but the inherent limitations of hand-crafted features led to poor localization results. With the rapid development of deep learning technology, various types of CVGL methods represented by convolutional neural networks and vision Transformer dominate the current research field. Deep learning-based methods [6], [7], [8], [9], [10], [11], [12], [13] focus on learning discriminative image representations to optimize the distance of image feature in feature space, and the core of these methods lies in a robust and efficient feature representation method. The image features obtained by this method can effectively deal with CVGL difficulties caused by viewpoint and scale differences.

Existing methods mainly improve the cross-view geo-localization by mining and exploiting the global contextual information of the image [6], [7], elaborating the backbone network for feature extraction [8], [9], introducing feature aggregation modules [10], and region alignment strategies [11]. Methods, such as attention mechanisms [12] and multiscale features [13], have also been widely used in CVGL. A large amount of work has shown that a robust and efficient feature representation method is effective and necessary in CVGL. Still, these methods are insufficient for mining the spatial location structure information in the image and do not fully consider the positional relationship information of the features and the redundancy of the information existing in the features.

To address the problems of existing methods, we propose a CVGL method that combines the global spatial relation attention (GSRA) and feature aggregation. The method uses GSRA to learn the spatial location structure information of features and uses the sinkhorn algorithm for locally aggregated descriptors (SALADs) to selectively discard useless information during the clustering and assignment of features to improve the generalization and robustness of descriptors. Our contributions are as follows.

1) The proposed GSRA can fully enhance the model's perception and discrimination ability by learning the information of features' spatial location structure. Due to the group convolution, GSRA has almost no effect on the complexity and memory occupancy of the model.
2) We integrate SALAD into CVGL, represent the aggregation of local features as an optimal transport problem and selectively discard useless information during the clustering and assignment of features to enhance the robustness and generalization of the descriptors.
3) Experiments have demonstrated the effectiveness and superiority of our method, and the proposed method achieves state-of-the-art (SOTA) results on the public University-1652 dataset.

Received 14 January 2025; accepted 17 February 2025. Date of publication 20 February 2025; date of current version 7 March 2025. This work was supported in part by the Joint Fund of Collaborative Innovation Center of Geo-Information Technology for Smart Central Plains, Henan Province, and in part by the Key Laboratory of Spatiotemporal Perception and Intelligent Processing, Ministry of Natural Resources under Grant 232109. *(Corresponding author: Yang Zhou.)*

The authors are with the Institute of Geospatial Information, Information Engineering University, Zhengzhou 450001, China (e-mail: zhouyang3d@163.com).

Digital Object Identifier 10.1109/LGRS.2025.3543949

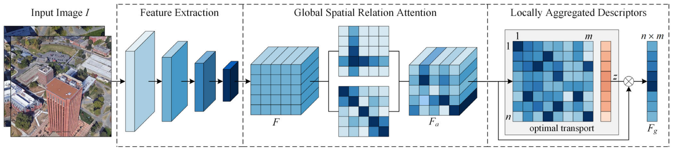

Fig. 1. Overview of the architecture for CVGL.

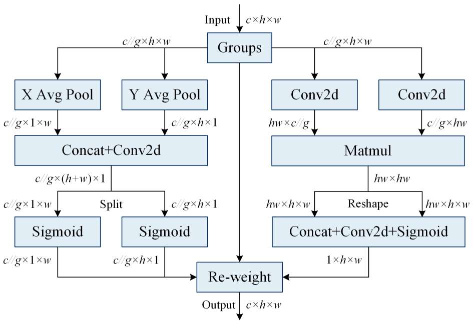

Fig. 2. Demonstration of GSRA.

II. METHODOLOGY

Our proposed framework is shown in Fig. 1. First, we extract local features using a ConvNeXt [14] network for a given input image $I \in R^{3 \times h_0 \times w_0}$. We cropped the last two layers of ConvNeXt to maximize the retention of the original model performance while drastically reducing the computation of the model. After the ConvNeXt network, we obtain a local feature $F \in R^{c \times h \times w}$, where $h \times w$ is the feature map size, and c is the number of channels. Then, we propose a GSRA module to enhance the model's contextual inference by learning the spatial location and structural information representation of features. Finally, we introduce SALAD, which aggregates the attention-weighted features F_a into global features F_g that represent the whole image to enhance the robustness and generalization of the representation.

A. Global Spatial Relation Attention

To make the model learn the viewpoint invariant features in the image more intensively and further utilize the feature location information as well as the relative positional relationship between the features to enhance the robustness of the mode. We propose a lightweight GSRA based on coordinate attention (CA) [15]. The proposed GSRA module employs the group convolution, which fully enhances the model's perceptual and discriminative abilities while having little effect on the complexity and memory occupancy of the model.

The basic structure of GSRA is shown in Fig. 2. For the feature $F \in R^{c \times h \times w}$ acquired by ConvNeXt, we first divide F into g subfeatures in the channel dimension, $F = [F_1, F_2, \ldots, F_i, \ldots, F_g]$, $F_i \in R^{c//g \times h \times w}$. Then, we use average pooling in the horizontal and vertical directions for the spatial dimensions of the subfeatures F_i. The embedding vector z^h of each feature x in F_i at height h can be expressed as follows:

$$z^h(h) = \frac{1}{w} \sum_{0 \le j < w} x(h, j). \tag{1}$$

Similarly, the embedding vector z^w of feature x at width w is obtained as follows:

$$z^w(w) = \frac{1}{h} \sum_{0 \le i < h} x(i, w). \tag{2}$$

The average pooling of horizontal and vertical directions in space enables GSRA to capture the spatial position information of features in these two directions. Then, we splice the features z^h and z^w along the height direction of the feature map and feed the spliced features into a 1×1 convolution, enabling the horizontal and vertical position information to interact and thus capture each feature's absolute spatial position information in F_i. Finally, we also need to separate and reduce the output of the 1×1 convolution into horizontal and vertical feature embedding and use the sigmoid function to obtain the attention scores of the spatial location information.

However, similar to CA, although the above operation can obtain the spatial location information of features, it cannot learn the global structural relationships between image features and lacks global contextual reasoning capability. Therefore, the other part of GSRA pairs acquires the global structural relationships between features by calculating the relationship matrix. Specifically, for any two features x_i and x_j in subfeature $F_i \in R^{c//g \times h \times w}$, the spatial relationship $r_{i,j}$ between them can be expressed by the following equation:

$$r_{i,j} = \theta_s(x_i)^\top \varphi_s\left(x_j\right) \tag{3}$$

where θ_s and φ_s denote the two spatial information encoding functions consisting of 1×1 convolution, batch normalization, and ReLU nonlinear activation functions. The spatial relation matrix between all features in subfeature F_i is denoted by $R_s \in R^{hw \times hw}$

Also, due to the spatial information coding function, the global spatial relation descriptor r_i of feature x_i should be expressed as a bidirectional relation between x_i and x_j

$$r_i = \left(r_{i,j}, r_{j,i}\right) = [R_s(i, :), R_s(:, i)]. \tag{4}$$

This is done by splicing $R_s(i, :)$ and $R_s(:, i)$ in the channel dimension, and then using 1×1 convolution and sigmoid function to obtain the attention scores regarding the relative spatial relationship information of the input features. The attention-weighted feature F_{ia} for each group can be obtained by broadcasting all the attention scores obtained by GSRA to the subfeatures F_i. The total attention-weighted feature $F_a \in R^{c \times h \times w}$ of the input feature $F = [F_1, F_i, \ldots, F_g]$ is the respliced combination of $[F_{1a}, F_{ia}, \ldots, F_{ga}]$ in the channel dimension.

B. Sinkhorn Algorithm for Locally Aggregated Descriptors

To obtain more efficient and robust descriptor representations, we introduce the SALAD [16], which represents the aggregation of local features as an optimal transport problem and selectively discards useless information during the clustering and assignment of features to improve the generalization and robustness of descriptors. In addition, we have adapted the structure of SALAD to connect with our backbone network.

The feature descriptor F_a obtained in the previous step is regarded as a set of local features $\{t_1, t_2, \ldots, t_i, \ldots, t_n\}$, where

$n = h \times w$, and t_i denotes a descriptor of length c. Then, we use a differentiable dimensionality reduction function $F(t)$ to downscale the input features, and the expression of the downscale function $F(t)$ is as follows:

$$F(t) = W_2(\sigma(W_1(t) + b_1)) + b_2 \tag{5}$$

where W_1 and W_2 denote the 1×1 convolution, σ denotes the ReLU nonlinear activation function, and b_1 and b_2 represent the induction bias.

After processing by the dimensionality reduction function $F(t)$, the dimension of the local feature t_i is changed from c to d. d is a hyperparameter controlling the length of the descriptor after clustering, and the value of d is 64 in this study.

We then use a score function $S(t)$ to compute the initial score matrix $S \in R^{n \times m}$ of the local feature t_i, where m is a hyperparameter denoting the number of clustering categories C, which is taken to be 128 in this study, and the elements $s_{i,j}$ in S denote the probability that the feature t_i is assigned to the jth clustering category C_j. The score function $S(t)$ consists of a set of learnable layers with the following expression:

$$S(t) = W_2(\sigma(W_1(t) + b_1)) + b_2. \tag{6}$$

We then use a learnable parameter z to assign uninformative features, thus assigning useless information to z during the learning process of the model. To do this, we need to expand the score matrix S to $\bar{S} \in R^{n \times (m+1)}$. Specifically,

$$\bar{S}_{i,m+1} = z\mathbf{1}_n \tag{7}$$

where $\mathbf{1}_n$ denotes a column vector of length n and value $\mathbf{1}$.

We reformulate the allocation of features as an optimal transmission problem, where the allocation of features $\mu = \mathbf{1}_n$ must be efficiently distributed among clusters C or sets $\kappa = [\mathbf{1}_m^\top, n - m]^\top$ of useless information. We use the sinkhorn algorithm to obtain the assignment matrix $\bar{P} \in P^{n \times (m+1)}$

$$\bar{P}\mathbf{1}_{m+1} = \mu, \quad \bar{P}^\top \mathbf{1}_n = \kappa. \tag{8}$$

Sinkhorn algorithm finds the optimal transport assignment between μ and κ by iteratively regularizing rows and columns from $\exp(\bar{S})$. Finally, it removes the useless information set to obtain the optimal assignment matrix $P = \bar{P}[:, :m]$. Then, the features are directly aggregated into their assigned clusters according to the assignment relationship in the optimal assignment matrix P

$$V_{j,k} = \sum_{i=1}^{n} P_{i,k} \cdot F\left(t_{i,k}\right) \tag{9}$$

where k represents the kth dimension of the descriptor t_i.

III. Experiments

A. Implementation Details

1) Dataset and Evaluation Metric: We experimented with and validated the proposed method using the University-1652, CVACT, and CVUSA dataset. University-1652 consists of 1652 outdoor scenes from 72 universities worldwide. The training set contains 701 scenes, and the test set includes 951 scenes, each scene containing one satellite image from Google Map and 54 drone images from Google Earth. The CVACT dataset contains panoramic-satellite image pairs of a portion of Canberra, Australia, and consists of 35 532 training and 8884 validation pairs. The CVUSA dataset was collected from the United States and contains 35 532 training and 8884 validation pairs.

As with most metrics used in CVGL methods, we use Recall@K (R@K) and average precision (AP) to evaluate the model's performance. R@K is more sensitive to the location of the first correctly matched image, while AP takes into account the locations of all correctly matched images and can more accurately reflect the comprehensive performance of the model. We use NVIDIA RTX A6000 for training and extract time per query is measured using NVIDIA RTX 3060 Laptop.

2) Hyperparameters: We optimized the model using semitriplet loss and AdamW optimizer. The initial learning rate is set to 0.0001, and we adopt a linearly decreasing learning rate adjustment strategy, where the learning rate is decayed every other batch until the last batch of training, where the learning rate will decay to 1/20 of the initial learning rate. For University-1652, we resized the dataset image to 384 × 384 for training and validation. For CVACT and CVUSA, the satellite images were processed using polar transform, and all images were resized to 256 × 512 before input to the network.

TABLE I

Comparison of Different Cross-View Geo-Localization Methods on University-1652

Methods	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
LPN	75.93	79.14	86.45	74.49
SAIG	78.85	81.62	86.45	78.78
TransFG	84.01	86.31	90.16	84.61
FSRA	85.50	87.53	89.73	84.94
MCCG	89.64	91.32	94.30	89.39
MFJR	91.87	93.15	95.29	91.51
Sample4Geo	92.65	93.84	95.14	91.39
Ours	**94.13**	**95.09**	**95.72**	**93.22**

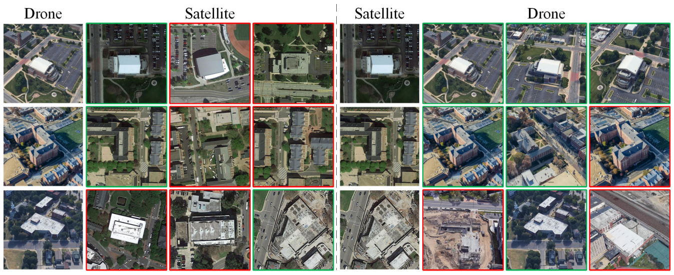

Fig. 3. Top five results of our method retrieval both in drone localization and drone navigation. The green boxes indicate the correct results.

B. Comparision With SOTAs

To fully illustrate the effectiveness of our proposed method, we compared our method with LPN [6], SAIG [8], TransFG [10], FSRA [11], MCCG [9], MFJR [7], and Sample4Geo [17] on the University-1652. The results of the experiment are presented in Table I.

In Table I, bold and underlined denote the optimal and suboptimal results of the experiments, respectively. Drone → satellite denotes the drone localization task in which the drone

TABLE II
COMPARISON OF DIFFERENT METHODS ON CVUSA AND CVACT

Methods	Params	Extract Time	CVUSA		CVACT	
	MB	ms	R@1	R@5	R@1	R@5
SAFA	2.8	20.2	89.84	96.93	81.03	92.80
TransGeo	22.4	21.5	94.08	98.36	84.95	94.14
SAIG	16.3	21.6	95.00	98.66	84.71	95.25
GeoDTR	21.3	22.4	95.43	98.86	86.21	**95.44**
Ours	13.1	26.1	**97.25**	**99.39**	**88.35**	95.32

image is employed as a query image, and satellite → drone represents the drone navigation task in which the drone image is employed as a reference image. In drone localization, our method achieves 94.13% and 95.09% of the best $R@1$ and AP, which are 1.48% and 1.25% higher than the suboptimal Sample4Geo. As for drone navigation, our method achieves 95.72% and 93.22% optimal $R@1$ and AP, which is 0.43% and 1.71% higher than the suboptimal method. The higher average accuracy of our method with close $R@1$ indicates the better overall performance of our architecture. In addition, our method has more minor differences in the two-way drone localization and navigation task, with a difference of 1.59% for $R@1$ and 1.87% for AP, which is significantly better than other methods. This indicates that the extracted feature representations of our method have better stability and can effectively discover similar contents in drone and satellite images to achieve better localization. Fig. 3 shows the experimental results for some of the images in the dataset. It is shown that our method performs well in most cases. However, our method cannot fully identify the correct results when the viewpoint, scale, and appearance are too different.

In addition, we compared with SAFA [18], SAIG [8], TransGeo [12], and GeoDTR [19] on CVACT and CVUSA datasets. The experimental results are shown in Table II. On the CVUSA and CVACT datasets, our method achieves the best $R@1$ of 97.25% and 88.35%, 1.82%, and 2.14% higher than the suboptimal method, respectively. Although the $R@5$ of our method on CVACT is slightly lower than that of the GeoDTR method, $R@1$ is significantly higher than that of the GeoDTR method, and the improvement is more excellent. Also, in CVGL, we pay more attention to $R@1$. Of course, our method also has shortcomings in terms of time efficiency, and we analyze the specific reasons in Section III-D.

C. Comparison of Different Attention Modules

We compared GSRA with CBAM [20], CA [15], ECA [21], and EMA [22] to illustrate the superiority of our proposed module. The experimental results are shown in Table III, where the baseline indicates that ConvNeXt is used as the feature extraction network, and NetVLAD [23] is used for feature aggregation without adding any attention. The experimental results show that, except for CBAM, adding the attention module on top of the baseline can effectively improve the model performance, and the effect of using CA, ECA, EMA, and GSRA becomes better in order. Our method has the most enormous improvement among several types of attention we compared. In drone localization, the $R@1$ and AP of our method are improved by 10.3% and 9.64%, respectively, which are 6.44% and 5.53% higher than suboptimal EMA. In drone navigation, $R@1$ and AP of our method are improved by 6.71% and 8.72%, respectively, which are 4.01% and 4.72% higher than that of EMA. CBAM, the only one among all methods that does not pay attention to spatial location information, is also the only one that shows accuracy degradation, illustrating the importance of learning spatial location information for CVGL. The above experiments provide the definitive evidence of the superiority of our proposed GSRA, which considers the positional information of the features in the space and the relative positional relationships among the features.

TABLE III
COMPARISON OF DIFFERENT ATTENTIONS

Methods	Drone→Satellite		Satellite→Drone	
	R@1	AP	R@1	AP
baseline	61.99	66.47	74.75	61.93
+CBAM	61.86	66.78	74.75	61.27
+CA	63.27	68.08	75.32	62.76
+ECA	64.78	69.71	77.03	64.01
+EMA	65.85	70.58	77.45	65.93
+GSRA	72.29	76.11	81.46	70.65

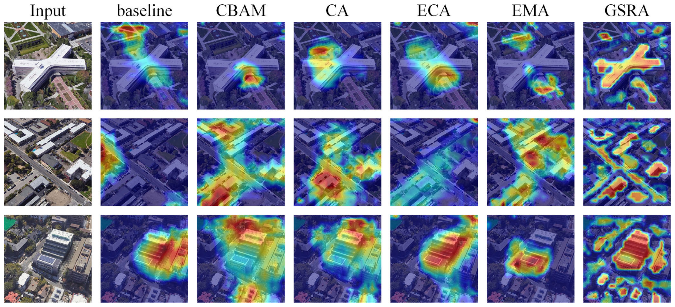

Fig. 4. Comparison of heat map visualizations for different attentions.

To more intuitively illustrate the advantages of GSRA over other attentions, we conducted experiments using GradCAM to visualize feature maps for models with different attentions added. The results of the experiments are presented in Fig. 4, where darker colors indicate that the model pays more attention to this region. Compared with other attentions, our approach is able to cover regions that are more favorable to CVGL. GSRA can reduce the attention to useless information regions, thus accurately highlighting the critical content in the image. In addition, the attention of the comparison methods is mainly focused on a specific region of the image, while our method can respond adequately to the other content regions of the image by taking into account the relative positional relationship between the features, thus mining more information that is beneficial for CVGL.

D. Ablation Study

To illustrate the effectiveness of each component in our approach, we performed ablation experiments on University-1652, and the results of the experiments are presented in Table IV. In the drone localization, after adding the proposed GSRA on the baseline, $R@1$ and AP are improved by 10.3% and 9.64%, respectively. After replacing NetVLAD in baseline with our introduced SALAD, $R@1$ and AP have

TABLE IV

Ablation Experiments on the Effectiveness of Model Components

Methods	Extract Time (ms)	Drone→Satellite		Satellite→Drone	
		R@1	AP	R@1	AP
baseline	21.7	61.99	66.47	74.75	61.93
+GSRA	21.9	72.29	76.11	81.46	70.65
+SALAD	25.3	93.29	94.36	95.72	92.70
+GSRA+SALAD	26.5	94.13	95.09	95.72	93.22

been improved by 31.3% and 27.89%, respectively. When we use GSRA and SALAD simultaneously, $R@1$ and AP have been enhanced by 32.14% and 28.62%, respectively, and the effect of using GSRA and SALAD simultaneously is better than using either alone. Similar results are also shown in drone navigation, in which after adding the GSRA module again on top of SALAD, $R@1$ is not improved, but AP is improved by 0.52%. It shows that the model maintains higher accuracy in a broader range of retrieval results and is able to rank the matched images more effectively, with better overall performance of the model.

Regarding time efficiency, the increase in inference time mainly comes from SALAD. This is because SALAD is an iterative process in computing the score matrix S, which significantly increases inference time. In the future, we will also focus on lightweight CVGL methods.

IV. Conclusion

We propose a CVGL method that combines GSRA with feature aggregation. First, we utilize the lightweight GSRA to learn the spatial location structure information of features, which fully enhances the model's perceptual and discriminative capabilities while having little effect on the number and complexity of the model's parameters. In addition, we introduce SALAD, which represents the aggregation of local features as an optimal transmission problem and selectively discards useless information during the clustering and assignment of features, thus improving the generalization and robustness of descriptors. Our method achieves SOTA results on public University-1652 with 94.13% and 95.09% of the best $R@1$ and AP in drone localization. Our method achieves 95.72% and 93.22% optimal $R@1$ and AP in drone navigation. Our proposed method achieved the best $R@1$ of 88.35% and 97.25% on the CVACT and CVUSA datasets. The experimental results demonstrate that our method performs excellently in learning more discriminative and robust visual representations, leading to better CVGL results. Our method has some shortcomings in terms of time cost, and we will focus on implementing lightweight methods to improve the CVGL's time efficiency in the future.

References

[1] X. Zhang, L. Wang, and Y. Su, "Visual place recognition: A survey from deep learning perspective," *Pattern Recognit.*, vol. 113, May 2021, Art. no. 107760.

[2] S. Ansari, F. Naghdy, and H. Du, "Human–machine shared driving: Challenges and future directions," *IEEE Trans. Intell. Vehicles*, vol. 7, no. 3, pp. 499–519, Sep. 2022.

[3] 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.

[4] 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.

[5] A. Viswanathan, B. R. Pires, and D. Huber, "Vision based robot localization by ground to satellite matching in GPS-denied situations," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 192–198.

[6] T. Wang et al., "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.

[7] F. Ge et al., "Multilevel feedback joint representation learning network based on adaptive area elimination for cross-view geo-localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5913915, doi: 10.1109/TGRS.2024.3396330.

[8] Y. Zhu, H. Yang, Y. Lu, and Q. Huang, "Simple, effective and general: A new backbone for cross-view image geo-localization," 2023, *arXiv:2302.01572*.

[9] T. Shen, Y. Wei, L. Kang, S. Wan, and Y.-H. Yang, "MCCG: A ConvNeXt-based multiple-classifier method for cross-view geo-localization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1456–1468, Mar. 2024.

[10] H. Zhao, K. Ren, T. Yue, C. Zhang, and S. Yuan, "TransFG: A cross-view geo-localization of satellite and UAVs imagery pipeline using transformer-based feature aggregation and gradient guidance," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 4700912, doi: 10.1109/TGRS.2024.3352418.

[11] 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.

[12] 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.

[13] F. Ge et al., "Multibranch joint representation learning based on information fusion strategy for cross-view geo-localization," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, 2024, Art. no. 5909516, doi: 10.1109/TGRS.2024.3378453.

[14] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, Jun. 2022, pp. 11966–11976.

[15] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13708–13717.

[16] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 17658–17668.

[17] F. Deuser, K. Habel, and N. Oswald, "Sample4Geo: Hard negative sampling for cross-view geo-localisation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 16801–16810.

[18] Y. Shi, L. Liu, X. Yu, and H. Li, "Spatial-aware feature aggregation for cross-view image based geo-localization," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, Dec. 2019, pp. 10090–10100.

[19] X. Zhang et al., "GeoDTR+: Toward generic cross-view geolocalization via geometric disentanglement," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10419–10433, Dec. 2024.

[20] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 3–19.

[21] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[22] D. Ouyang et al., "Efficient multi-scale attention module with cross-spatial learning," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2023, pp. 1–5.

[23] 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.