Cross-Attention Between Satellite and Ground Views for Enhanced Fine-Grained Robot Geo-Localization

Dong Yuan[1], Frederic Maire[1], Feras Dayoub[2]

[1]QUT Centre for Robotics, Queensland University of Technology, Australia {yuand2, f.maire}@qut.edu.au

[2]Australian Institute for Machine Learning (AIML), University of Adelaide, Australia feras.dayoub@adelaide.edu.au

Abstract

Cross-view image geo-localization aims to determine the locations of outdoor robots by mapping current street-view images with GPS-tagged satellite image patches. Recent works have attained a remarkable level of accuracy in identifying which satellite patches the robot is in, where the location of the central pixel within the matched satellite patch is used as the robot coarse location estimation. This work focuses on robot fine-grained localization within a known satellite patch. Existing fine-grain localization work utilizes correlation operation to obtain similarity between satellite image local descriptors and street-view global descriptors. The correlation operation based on liner matching simplifies the interaction process between two views, leading to a large distance error and affecting model generalization. To address this issue, we devise a cross-view feature fusion network with self-attention and cross-attention layers to replace correlation operation. Additionally, we combine classification and regression prediction to further decrease location distance error. Experiments show that our novel network architecture outperforms the state-of-the-art, exhibiting better generalization capabilities in unseen areas. Specifically, our method reduces the median localization distance error by 43% and 50% respectively in the same area and unseen areas on the VIGOR benchmark.

1. Introduction

Cross-view image geo-localization has proven to improve outdoor robot localization accuracy in environments with noisy GPS signals [2, 25]. Previous works formulate the cross-view geo-localization problem as image retrieval, matching current street-view images with GPS-tagged satellite patches in a reference database [3, 8, 17, 20–22, 30, 32]. The GPS coordinate corresponding to the central pixel within the retrieved satellite patch is used as the current coarse location estimation. Even with high image retrieval accuracy, the coarse location estimation can potentially result in tens of meters errors, due to the possibility of street view images being captured at a significant distance from the center of the satellite image [33], see Figure 1. In this work, we focus on fine-grained localization within a known satellite image patch, *i.e.* to predict satellite image pixel coordinates corresponding to street-view images.

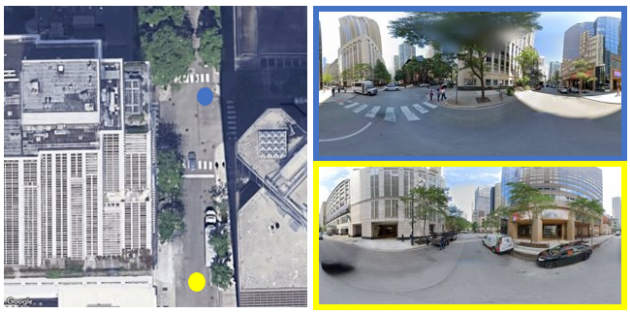


Figure 1. An example of ground-aerial view image pairs. Two Street View images are covered by a satellite image, and the two street View images are located far away from the center of the satellite image.

Zhu *et al*. [33] first attempts to address the fine-grained localization problem on the VIGOR benchmark by offset regression prediction. The regression prediction header is simply built on the top of the connection of street-view and satellite image global descriptors, which ignores the correlation between the local features of the two views. Xia *et al*. [28] formulated the fine-grained localization as a multi-class classification problem. Their model encodes satellite and street-view images into local and global descriptors respectively and calculates the correlation between these descriptors to produce similarity score maps. The similarity score maps are up-sampled to obtain probability heat maps of the same size of the input satellite images. which are employed for coordinate prediction. However, correlation operation as a linear matching will lose details and the up-sampling process may also bring unexpected noise.

To address these limitations, inspired by the core idea of Transformer [24], we design a cross-view feature fusion network, which mainly consists of self-attention layers and cross-attention layers. Compared to correlation operation in [28], cross-attention layers facilitate interaction between local features from two different views and effectively integrate semantic and spatial information. To make the localization more accurate, we introduce a prediction network built on top of the cross-view fusion features, which consists of a classification header and a regression header. The classification header predicts an index indicating the location of a grid area within the satellite image and the regression header predicts an offset value based on the classification results. Our main contributions can be summarized as follows:

- We propose a novel Transformer-based cross-view feature fusion network and leverage the cross-attention mechanism to present the cross-correlation between aerial and street views. Compared with descriptor similarity matching, our cross-attention module can exploit mutual interactions between local features of two views to establish more complex correspondences.
- We for the first time combine classification prediction and regression prediction to address the cross-view fine-grained localization problem and validate that this coarse-to-fine prediction manner can efficiently further improve localization accuracy.
- Our experimental results on the VIGOR benchmark show that the proposed cross-view fine-grained localization framework significantly outperforms the state-of-the-art method in predicting the street-view location which is away from the center of the satellite image. Compared with the state-of-the-art, our model achieves lower prediction distance errors in cross areas, thereby indicating better generalization.

To foster future research on fine-grained robot localization, we make our code available at: https://github.com/UQ-DongYuan/CVLocationTrans

2. Related work

Cross-view Image Retrieval. Cross-view image retrieval task is to find the corresponding satellite image patch from a reference database to the current ground view image, thereby using the GPS tag of the matched satellite patch as the current localization estimate. To explore this task, [31] and [11] proposed two large-scale ground-to-satellite image benchmarks, namely CVUSA and CVACT. In each image pair, a ground-view image corresponds to a satellite image center, and the orientations of the two views are aligned. This alignment ensures that the top part of the satellite images (representing the North direction) matches with the center of the street-view images. Recent works [3, 8, 17, 20, 22, 30, 32] explore improving image retrieval accuracy on these two benchmarks. CVM-Net [8] applies NetVLAD [1] to generate view-invariant descriptors for cross-view image pairs, enhancing matching accuracy. SAFA [20] introduces a spatial attention module to establish spatial associations across views. To bridge domain gap between the ground and aerial views, some works [13, 16, 17] adopt conditional Generative Adversarial Networks (cGANs) [9] to synthesise one view from another by utilizing depth or semantic information, other works [20, 21] geometrically transform aerial view to ground view via the association between azimuth directions in the satellite image and vertical lines in the street-view image. Roughly eliminating the domain gap between two views eases the descriptor learning process and improves the cross-view image retrieval accuracy. Advanced Transformer-based architectures have also been explored in [30, 32], and the positional encoding of the Transformer is helpful in integrating spatial correspondence between two views. Cross-view image retrieval methods can effectively find the corresponding satellite image patch, but it is not practical for real-world applications as the location of the street-view image is not invariably at the center of the satellite image patch.

Cross-view Fine-grained Localization. To break one-to-one center-alignment correspondence, [33] proposed a novel benchmark called VIGOR, in which every ground-view image is covered by four distinct satellite patches. The ground-view images are captured at arbitrary locations within their corresponding satellite patches. [33] applies the same module in [20] to build global descriptors of ground-view and satellite images. The extracted descriptors from two views are fused to regress the offset between the ground image and the satellite patch center. [28] employs a Siamese-like network to encode a pair of ground-view and satellite images into a global descriptor and $N \times N$ local descriptors, respectively. The generated descriptors are employed to create an $N \times N$ similarity score map through a correlation operation. Subsequently, following a U-Net-like up-sampling process [18], the similarity score map is up-sampled into a high-resolution heat map, which signifies the probability of the ground-view location. However, correlation operation as a linear matching process cannot effectively integrate complicated non-linear correspondences between ground and aerial views, while the subsequent up-sampling process can also contribute to prediction errors.

Transformer and Cross Attention. Transformer was first proposed in [24] and has been widely applied in natural language processing (NLP). Dosovitskiy *et al*. [5] were the first to introduce the Transformer architecture in computer vision tasks and applied it to image classification. DETR [4]

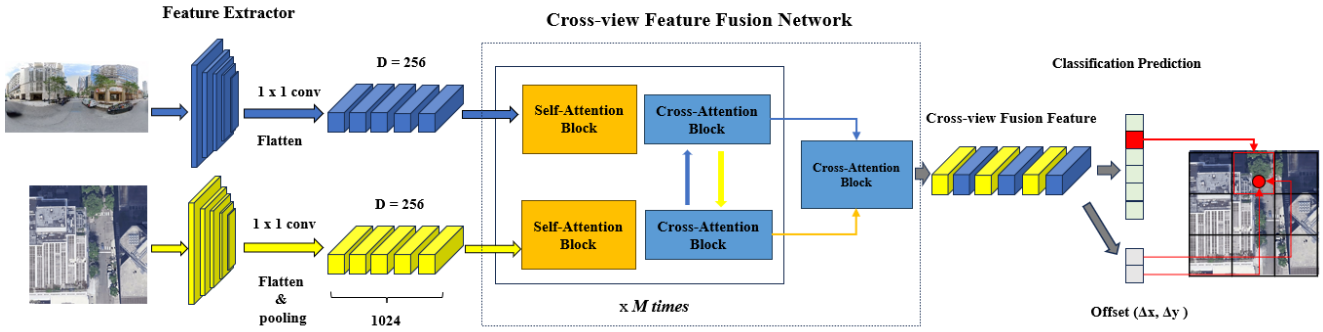


Figure 2. Overview of the proposed architecture. The architecture mainly consists of three components: local feature extractor, cross-view feature fusion network and prediction headers. Two local feature extractors separately extract features from ground-view and satellite images.

follows the Encoder-Decoder architecture of Transformer to achieve end-to-end objection detection. Recently, Transformers have also been employed in the semantic segmentation task and achieved state-of-the-art results [23, 29]. As the core idea of the Transformer decoder, the cross-attention mechanism can establish correlations between input and output sequences of varying lengths, thereby achieving the objective of sequence-to-sequence transformation. Motivated by this, we argue that utilizing the cross-attention mechanism instead of the correlation operation can better integrate the correspondences between ground and aerial views. Consequently, the fused features following the cross-attention process are more informative and can be promising to improve localization accuracy and generalization ability.

3. Methodology

This section introduces the proposed Transformer-based method. Our proposed architecture consists of three main components: local feature extractor, cross-view feature fusion network and prediction header. The local feature extractor is a Siamese-like network for extracting local features from ground and aerial views separately. The extracted features as input are passed through the cross-view feature fusion network to establish correspondences between two views and obtain the cross-view fusion features. Finally, a classification header and a regression header are built on top of the fusion features for coarse-to-fine location prediction. Further details of each components will be illustrated in the subsequent parts and an overview of the proposed method is presented in Figure 2.

3.1. Local Feature Extraction

We apply two modified ResNet50 [7] to extract features from ground-view and satellite images separately. Specifically, we employ the convolutional layers from the first four stages to extract features, change the downsampling stride of the fourth stage to 1, and modify the stride of the dilation convolution to 2. These adjustments aim to preserve the high-resolution feature map while concurrently enlarging the receptive field. For the ground-view images G, feature maps $f_g \in \mathbb{R}^{C \times H_g \times W_g}$ are obtained after passing through the feature extractor, where H_g and W_g are $1/8$ of the height and width of the input ground-view image. For the satellite images S, to ensure a consistent feature map resolution for classification prediction, we incorporate an adaptive pooling process at the end of the feature extractor. Finally, the branch of satellite image feature extraction outputs feature maps $f_s \in \mathbb{R}^{C \times N \times N}$. The channel dimension C will be reduced to D using a 1×1 convolution layer before feature fusion.

3.2. Cross-view Feature Fusion Network

We first briefly introduce the cross-view feature fusion process and further details of attention mechanisms will be presented later. Our proposed cross-view feature fusion network mainly comprises self-attention blocks and a cross-attention blocks, employed for enhancing and fusing image features, respectively. Specifically, the ground and aerial view feature maps from the feature extraction stage are further flattened, obtaining feature representations $f_{g'} \in \mathbb{R}^{D \times H_g W_g}$ and $f_{s'} \in \mathbb{R}^{D \times N^2}$. As depicted in Figure 2, these two feature representations are first fed into two separate self-attention blocks (SAB) for feature enhancement. Subsequently, two followed cross-attention blocks (CAB) receive feature information from both ground and aerial views, facilitating correlation integration and feature fusion. In this manner, the feature fusion module comprising two SABs and two CABs will be repeated M times, and another CAB will be utilized in the end to obtain the ultimate cross-view fusion feature $f_{fusion} \in \mathbb{R}^{D \times N^2}$.

Multi-head Attention. The key component of the

Transformer is the multi-head attention block. The input feature representations are converted into three linear projections with dimension d, conventionally named query (Q), key (K) and value (V), as the input of the attention layers. The attention operation is denoted as:

$$\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

Intuitively, the query Q retrieves information from the value V based on the attention weight. which is computed from the scale dot-product of Q and the key K corresponding to each value V. In practice, each Q, K and V is typically divided into k different heads, and all the k heads participate in attention operation in parallel, as these multiple heads consider information from different representation subspaces at different positions [24].

Positional Encoding. Positional encoding is added to maintain the positional information. Following DETR [4], we apply fixed 2D positional encodings generated by a sine function. In contrast to ViT [5], which adds positional encodings to the backbone output only once, in this work, the fixed positional encodings are added to queries and keys in every self-attention and cross-attention block, as illustrated in Figure 3. We refer readers to the supplementary material for more details of positional encoding.

Cross-attention Blocks. The core idea of our architecture is to employ the cross-attention mechanism to establish correspondence between two views. The details of the cross-attention block are shown in Figure 3. The cross-attention block receives feature representations, $f_{g'}$ and $f_{s'}$, from the ground and aerial views, respectively, in order to create Query (Q), Key (K) and Value (V):

$$\begin{aligned} Q_g &= W_{Q_g} \times f_{g'} \\ K_s &= W_{K_s} \times f_{s'} \\ V_s &= W_{V_s} \times f_{s'}, \end{aligned} \tag{2}$$

where Query is projected from the ground view, Key and Value are from the aerial view. W_{Q_g}, W_{K_s} and W_{V_s} are weight matrices to be learned. The attention score can be computed between the query and the key:

$$\text{Scores} = Q_g^T \times K_s. \tag{3}$$

Then, the obtained attention score is normalized using Softmax function:

$$\text{SoftmaxScores} = \text{Softmax}(\text{Scores}) \tag{4}$$

Finally, we obtain the new ground-view feature representations, $f'_{g'}$, by taking the weighted sum of the values vectors:

$$f'_{g'} = \text{SoftmaxScores} \times V_s \tag{5}$$

As a result, $f'_{g'}$ will be the new ground-view feature representations with cross-attention applied to them from the satellite-view feature representations $f_{s'}$. In essence, the cross-attention mechanism enables the model to weigh the importance of different parts of $f_{s'}$ when updating $f_{g'}$, thereby fusing relevant information from both the satellite and gournd views.

Furthermore, a feed-forward network (FFN) module consisting of two linear projection layers with $ReLU$ activation function in between is employed to further enhance cross-attention features.

3.3. Multi-class Classification

We build a multi-class classification header on top of the fusion features f_{fusion}, which is a three-layer perceptron with hidden dimension d and $ReLU$ activation function. The output of the classification header is a vector with dimension $N^2 \times 1$, indicating that there are N^2 total classification categories. As shown in Figure 4a, in our classification task, the input satellite image is divided into $N \times N$ grid areas, and the classification header aims to predict the grid in which the current street-view image is located. We utilize the standard categorical cross-entropy loss for training our classification header, which is defined as:

$$\mathcal{L}_{ce} = -\sum_{i}^{C} y_i \log(\text{softmax}(p_i)), \tag{6}$$

where y_i denotes the ground-truth label and p_i denotes the output of the classification header.

3.4. Coordinate Offset Regression

The offset regression header is another branch based on the fusion features, of which the perception layers are similar to the classification header. Unlike the classification header, the regression output vector has dimension $N^2 \times 2$, intended for coordinate offset regression. In [33], the proposed model aims to regress the offset between the street-view location and the center of the satellite image. Despite the offset being normalized with the size of the satellite image during training, the wide prediction range still poses a challenge for accurate predictions. In contrast, our proposed offset regression is based on the classification results. Specifically, as shown in Figure 4b, the coordinates of any point (g_x, g_y) in the $N \times N$ grids can be defined as:

$$\begin{aligned} g_x &= \sigma(t_x) + c_x \\ g_y &= \sigma(t_y) + c_y, \end{aligned} \tag{7}$$

where (c_x, c_y) is the coordinate corresponding to the top left corner of the grid, and (t_x, t_y) is the offset relative to the top left corner of the grid. σ is the $Sigmoid$ activation function, which limits the predicted offset between 0 and 1.

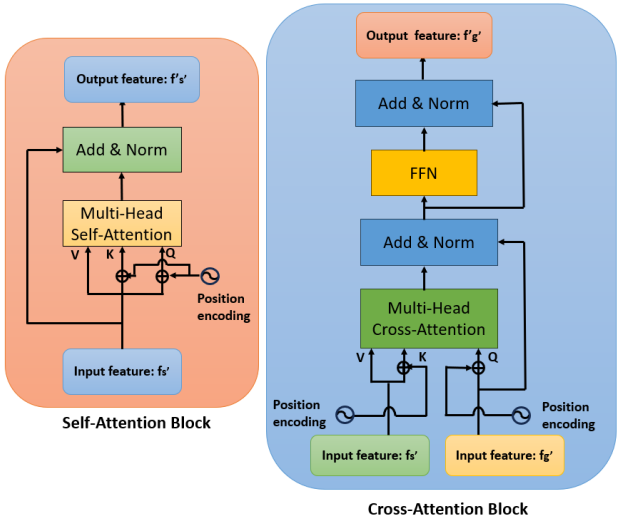
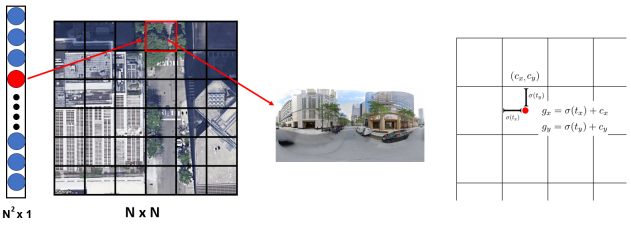


Figure 3. Left: Self-attention Block (SAB). Right: Cross-attention Block (CAB). The cross-attention block receives features from two views. One view feature provides value V and key K, another view feature provides query Q. The positional encoding is added to Q and K to preserve spatial information.



(a) Illustration of multi-class classification prediction

(b) Illustration of offset regression.

Figure 4. Illustration of the coarse-to-fine localization process. Our approach first predicts the located grid index in the $N \times N$ grid map and regresses offset based on the classification results.

Consequently, our regression header aims to predict the offset t_x and t_y related to the grid selected by the classification header. We employ the mean square error to train the offset regression header. The regression loss can be defined as:

$$\mathcal{L}_{mse} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{8}$$

where y_i denotes the ground-truth offset and $\hat{y}_i$ denotes the prediction offset values.

3.5. Model Optimization

Since each attention block requires new queries (Q), keys (K), and values (V) to participate in the calculations, the number of parameters in our model will be higher than that of the convolutional neural networks (CNNs)-based model. This leads to a concern that the model might be more prone to overfitting. In addition, for each ground-aerial image pair, the orientations are aligned. Therefore, simple data augmentation techniques like image rotation or flipping cannot be performed. To address this issue, we utilize the Adaptive Sharpness-Aware Minimization (ASAM) [10] optimization method, which is also employed in [32] for training transformer-based model. Specifically, ASAM function is seeking parameters that lie in neighborhoods having uniformly low loss value, rather than focusing on paramters that only themselves have low loss value. Consequently, ASAM can simultaneously minimize loss value and loss sharpness to overcome the overfitting issue. We refer readers to the supplementary material for more detailed descriptions.

4. Experiments

4.1. Datasets and Evaluation Metrics

We conduct experiments on two benchmarks, namely VIGOR [33] and Oxford RobotCar [14, 15], to evaluate the ability of cross-view fine-grained localization. We choose the state-of-the-art multi-class classification-based method (MCC) [28] as our baseline and compare all experimental results on both benchmarks.

VIGOR Dataset. VIGOR contains 238 696 ground-view panoramas and 90 618 aerial images from four cities, *i.e*. New York City (Manhattan), San Francisco, Chicago, and Seattle. As defined in [33], each ground-view panorama corresponds to 1 positive and 3 semi-positive aerial images. If an aerial image is positive, it indicates that the ground-view panorama is captured within the central quarter area of the aerial image, otherwise, it is semi-positive. As a result, ground-view locations could be placed arbitrarily within a satellite image, not just at the center. [33] also defined two training-testing protocols, namely the 'same-area' and 'cross-area' protocols. In the same-area protocol, all ground-aerial images from the four cities are used for both training and testing. Conversely, in the cross-area protocol, images from New York and Seattle are used for training, while images from the other two cities are used for testing. The cross-area protocol introduces more challenges to cross-view fine-grained localization task and enables the evaluation of the model's generalization ability. Furthermore, each aerial image with GPS tags corresponds to a ground resolution of 0.114 meters, which can be employed for meter-level evaluation. We conduct experiments using both positive and semi-positive aerial images and adopt same-area and cross-area protocols for training and testing.

Oxford RobotCar Dataset. The Oxford RobotCar dataset provides multiple sensor data, including 72 traversals of a route through Oxford under different illumination, weather, and traffic conditions [15]. Following the data split

Model	Same-Area				Cross-Area			
	Positives		Pos+Semi-Pos		Positives		Pos+Semi-Pos	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
VIGOR [33]	10.55	9.31	16.64	13.82	11.26	10.02	18.66	16.73
MCC [28]	9.86	4.58	13.45	5.39	13.06	6.31	17.13	7.78
Ours	**6.72**	**2.68**	**9.32**	**3.11**	**7.28**	**2.83**	**9.78**	**3.89**

Table 1. Localization errors on VIGOR [33]. MCC [28] is the baseline method. Best performance in bold.

in MCC [28], we have 17067, 1698 and 5089 ground-level front view images in the training, validation and test sets respectively. The test set consists of three different traversals. MCC [28] stitches all corresponding satellite image patches provided by [26,27] together to create a continuous satellite map. As described in [28], during training, satellite image patches are randomly sampled from the continuous satellite map around the ground image locations. The orientations of each ground-satellite image pair are aligned, and the resolution of each satellite image patch is 800 $pixels$ $\times$ 800 $pixels$ corresponding to 73.92m $\times$ 79.92m on the ground.

Evaluation Metrics. Following the baseline method MCC [28], we utilize the mean and median distance errors, measured in meters, between the ground truth and predicted locations for model evaluation. For the Oxford RobotCar test set, we first obtain the mean and median distance errors for each of the three distinct traversals individually, and subsequently calculate the average and standard deviation across the three tests.

4.2. Implementation Details

For the VIGOR dataset, ground-view panorama and satellite images are resized to 256 $\times$ 512 and 256 $\times$ 256 respectively during both training and testing. For the Oxford RobotCar dataset, front-view and satellite images are resized to 256×384 and 512×512 respectively. The aerial-view feature map dimension $C \times N \times N$ is $1024 \times 32 \times 32$ and the channel dimension $C = 1024$ will reduced to $D = 256$ by 1×1 convolutional layers. The feature fusion module will be repeated 4 times and the hidden dimension d of prediction headers is 256. The local feature extractor parameters are initialized with ResNet50 [7] pre-trained weights on ImageNet [19], and other parameters are initialized with Xavier init [6]. We employ AdamW [12] for model optimization and set feature extractor's learning rate to 10^{-5}, other parameters' learning rate to 10^{-4}.

4.3. VIGOR Same-area Generalization

Our proposed architecture is trained and tested following the VIGOR same-area protocol. A summary of the experimental results comparison between our approach and the baseline MCC [28] is presented in Table 1 Same-Area. Following the baseline MCC [28], we conduct two types of testings, namely Positives and Pos+Semi-pos, as shown in Table 1. For the Positive testing, only positive satellite images are utilized for location prediction. In the case of positive satellite images, the ground-view location is situated near the center of the satellite image, which makes it comparatively easier for location prediction. The mean distance error of our approach and MCC [28] are both within 10 meters, and our approach has 32% and 41% lower mean and median localization error than MCC [28]. For the Pos+Semi-pos testing, all of positive and semi-positive satellite images are employed for location prediction. Predicting ground-view locations within a semi-positive satellite image is more challenging since some of the semantic information, such as buildings and cars, may not be present in the aerial view, especially when the ground-view location is near the edge of the satellite image. Compared to MCC [28], the mean distance error of our approach is still within 10 meters, and our approach has 31% and 42% lower mean and median localization error than MCC [28]. Improved Pos+Semi-pos testing results indicate that our cross-attention blocks can effectively establish correspondences between two views, even when the semantic information is not consistent in the two views.

4.4. VIGOR Cross-area Generalization

Predicting the location of a new street-view image in an unseen area is a more difficult task since street views look very different in different cities. As shown in Table 1 Cross-Area, our model generalizes well under this challenging setting in terms of Positive testing and Pos+Semi-pos testing. Compared to MCC [28], our cross-attention operates on the feature maps of the two views, facilitating greater interactions for the local features of both views. As a result, our approach has 43% and 50% lower mean and median localization error than MCC [28] in terms of the Pos+Semi-pos testing.

In addition, we observed that our model shows little gaps between the Same-area and Cross-area settings. Specifically, for the Positive+Semi-positive test, our model exhibits mean errors of (9.32 *vs.* 9.78) and median errors of (3.11 *vs.* 3.89) in the Same-area and Cross-area settings, respectively. In contrast, the performance of MCC [28] degenerates severely as the test setting becomes more chal-

Ablation	Same-Area				Cross-Area			
	Positives		Pos+Semi-Pos		Positives		Pos+Semi-Pos	
	Mean	Median	Mean	Median	Mean	Median	Mean	Median
Ours w/o ASAM	8.50	3.32	11.29	3.88	10.84	4.30	13.80	5.55
Ours	**6.72**	**2.68**	**9.32**	**3.11**	**7.28**	**2.83**	**9.78**	**3.89**

Table 2. Ablation study of ASAM on VIGOR [33]. Best performance in bold.

lenging. This demonstrates that our model exhibits enhanced robustness under difficult tests.

We argue that the Pos+Semi-pos testing under the Cross-Area setting is more representative of real-world scenarios for robot geo-localization. A robot should possess the capability to localize itself in an unfamiliar environment, where its location could be situated anywhere within the area covered by the available satellite image. Drawing from the aforementioned experimental results, our proposed approach holds greater promise for practical applications in robot geo-localization.

4.5. Oxford RobotCar Results and Analysis

Following MCC [28], we trained and tested the proposed model on the Oxford RobotCar benchmark, and the experimental results are detailed in Table 3. Our model achieved competitive results, specifically obtaining the same average mean error but with a lower standard deviation of the average. Analyzing the results, we believe that there are two reasons why our model's performance is affected. First, in comparison to ground-view panoramas, front-view images lack a significant amount of information. This leads to the situation where most local features of satellite images cannot be effectively associated with the features from front-view images in the cross-attention process. Another reason is that the number of images in the Oxford RobotCar dataset used for training is considerably smaller than that of VIGOR. Insufficient data can hinder the effective training of the transformer-based model, even when employing ASAM [10]. In future work, we can consider utilizing the Oxford RobotCar's multi-view street images to enhance the performance of the model.

4.6. Ablation Study

ASAM. We did an ablation study on the impact of using ASAM [10] on the model performance, as shown in Table 2. ASAM decreases all localization mean and median errors on the VIGOR dataset. In addition, 'Ours w/o ASAM' still outperforms the baseline MCC [28] by a large margin, especially under the cross-area setting. This indicates that the cross-attention-based feature fusion method holds a significant advantage over the correlation operation-based method.

Model	Mean	Median
VIGOR [33]	2.29±0.31	1.72±0.21
MCC [28]	1.77 ±0.25	**1.24±0.10**
Ours	**1.77±0.20**	1.32±0.08

Table 3. Experimental results on the Oxford RobotCar [15]. Shown are the average ± standard deviation of 'mean' and 'median' errors over 3 test traversals. Best results in bold.

5. Conclusions

In this work, we focus on cross-view fine-grained localization within a known satellite image. We propose employing cross-attention instead of correlation operations to establish correspondences between ground and aerial views. Additionally, we combine multi-class classification and offset regression to achieve accurate fine-grained localization. Our proposed method exhibits a significant performance improvement over the state-of-the-art approach on the VIGOR dataset, particularly in the more challenging setting (cross area). Competitive results are also achieved on the Oxford RobotCar dataset. This demonstrates that cross-attention-based cross-view feature fusion can enhance the robustness and generalization ability of the model. Future work will address ground-view image sequence input and fine-grained orientation estimation.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, pages 5297–5307, 2016. 2

[2] Mayank Bansal, Harpreet S Sawhney, Hui Cheng, and Kostas Daniilidis. Geo-localization of street views with aerial image databases. In *ACM MM*, pages 1125–1128, 2011. 1

[3] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *ICCV*, October 2019. 1, 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, pages 213–229. Springer, 2020. 2, 4

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*, 2021. 2, 4

[6] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010. 6

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3, 6

[8] Sixing Hu, Mengdan Feng, Rang MH Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *CVPR*, pages 7258–7267, 2018. 1, 2

[9] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2

[10] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. In *International Conference on Machine Learning*, pages 5905–5914. PMLR, 2021. 5, 7

[11] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *CVPR*, pages 5624–5633, 2019. 2

[12] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2018. 6

[13] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *CVPR*, pages 859–867, 2020. 2

[14] Will Maddern, Geoffrey Pascoe, Matthew Gadd, Dan Barnes, Brian Yeomans, and Paul Newman. Real-time kinematic ground truth for the oxford robotcar dataset. *arXiv preprint arXiv: 2002.10152*, 2020. 5

[15] Will Maddern, Geoff Pascoe, Chris Linegar, and Paul Newman. 1 Year, 1000km: The Oxford RobotCar Dataset. *The International Journal of Robotics Research (IJRR)*, 36(1):3–15, 2017. 5, 7

[16] Krishna Regmi and Ali Borji. Cross-view image synthesis using geometry-guided conditional gans. *Computer Vision and Image Understanding*, 187:102788, 2019. 2

[17] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *ICCV*, pages 470–479, 2019. 1, 2

[18] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 2

[19] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115:211–252, 2015. 6

[20] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *NeurIPS*, 32, 2019. 1, 2

[21] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *CVPR*, pages 4064–4072, 2020. 1, 2

[22] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *AAAI*, volume 34, pages 11990–11997, 2020. 1, 2

[23] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *ICCV*, pages 7262–7272, 2021. 3

[24] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 2, 4

[25] Anirudh Viswanathan, Bernardo R Pires, and Daniel Huber. Vision based robot localization by ground to satellite matching in gps-denied situations. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 192–198. IEEE, 2014. 1

[26] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Geographically local representation learning with a spatial prior for visual localization. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 557–573. Springer, 2020. 6

[27] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Cross-view matching for vehicle localization by learning geographically local representations. *IEEE Robotics and Automation Letters*, 6(3):5921–5928, 2021. 6

[28] Zimin Xia, Olaf Booij, Marco Manfredi, and Julian FP Kooij. Visual cross-view metric localization with dense uncertainty estimates. In *ECCV*, pages 90–106. Springer, 2022. 1, 2, 5, 6, 7

[29] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021. 3

[30] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *NeurIPS*, 34:29009–29020, 2021. 1, 2

[31] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *CVPR*, July 2017. 2

[32] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *CVPR*, pages 1162–1171, 2022. 1, 2, 5

[33] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *CVPR*, pages 3640–3649, 2021. 1, 2, 4, 5, 6, 7