Cross-view Geo-localization with Evolving Transformer

Anonymous Author(s)
Affiliation
Address
email

Abstract

In this work, we address the problem of cross-view geo-localization, which estimates the geospatial location of a street view image by matching it with a database of geo-tagged aerial images. The cross-view matching task is extremely challenging due to drastic appearance and geometry differences across views. Unlike existing methods that predominantly fall back on CNN, here we devise a novel evolving geo-localization Transformer (EgoTR) that utilizes the properties of self-attention in Transformer to model global dependencies, thus significantly decreasing visual ambiguities in cross-view geo-localization. We also exploit the positional encoding of Transformer to help the EgoTR understand and correspond geometric configurations between ground and aerial images. Compared to state-of-the-art methods that impose strong assumption on geometry knowledge, the EgoTR flexibly learns the positional embeddings through the training objective and hence becomes more practical in many real-world scenarios. Although Transformer is well suited to our task, its vanilla self-attention mechanism independently interacts within image patches in each layer, which overlooks correlations between layers. Instead, this paper propose a simple yet effective self-cross attention mechanism to improve the quality of learned representations. The self-cross attention models global dependencies between adjacent layers, which relates between image patches while modeling how features evolve in the previous layer. As a result, the proposed self-cross attention leads to more stable training, improves the generalization ability and encourages representations to keep evolving as the network goes deeper. Extensive experiments demonstrate that our EgoTR performs favorably against state-of-the-art methods on standard, fine-grained and cross-dataset cross-view geo-localization tasks. Codes are available in the supplementary material.

1 Introduction

Estimating the geospatial location of a given image is of paramount importance for robot navigation [11], 3D reconstruction [12] and autonomous driving [5]. Recently, cross-view geo-localization, which aims to match query ground images with geo-tagged database aerial/satellite images, has emerged as a promising proposal to address this problem. Despite its appealing application prospect, the cross-view matching task is extremely challenging due to drastic viewpoint changes between ground and aerial images. Thus, it is critical to understand and correspond both image content (appearance and semantics) and spatial layout across views.

Towards the above goal, several recent works incorporate convolutional neural networks (CNNs) with NetVlad layers [8], capsule networks [20] or attention mechanisms [2, 16] to learn visually discriminative representations. However, the locality assumption of their CNN architectures hinders their performance in complex scenarios, where visual interferences such as obstacles and transient objects (*e.g.*, cars and pedestrian) may exist. Instead, human visual system utilizes not only *local*

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

information but also *global* context to make more accurate predictions when visual signals are ambiguous or incomplete. Another branch of works exploits geometry prior knowledge to reduce ambiguities caused by geometric misalignments. Though promising, these methods either rely heavily on predefined orientation prior [9], or make a restrictive assumption that ground and aerial images are orientation-aligned [16]. As a result, such strong assumptions limit the applicability of these approaches, which prompts us to seek a more flexible approach for encoding position-aware representations.

Motivated by these observations, we introduce Transformer [21], which exceeds in global contextual reasoning and thus can be naturally employed to reduce visual ambiguities in cross-view geo-localization. Besides, the positional encoding of Transformer enables our network to flexibly learn position-dependent representations. Specifically, our proposed evolving geo-localization Transformer (EgoTR) is built upon two independent Vision Transformer (ViT) [4] branches, which split a feature map into several sub-patches while modeling interactions between arbitrary patches. We show in the experiment that due to its context- and position-dependent natures, such a Transformer-based network is a well suited candidate for cross-view geo-localization and shows its superiority compared to the dominant CNN-based counterparts.

We also take a deep look at self-attention map, which is the integral part of Transformer and is independently learned in each Transformer block. Nevertheless, such an independent learning strategy overlooks correlations between layers. Specifically, relating features from adjacent layers could improve the representation ability of network [13]. Compared to simply fusing features from all layers, interacting between features from adjacent layers produces less information redundancy and noise interference as they are in closer abstraction level. To explore cross-layer correlations, we replace the self-attention with a novel self-cross attention mechanism. Simple yet effective, the proposed self-cross attention learns pairwise similarities between features of adjacent blocks rather than of the same block. On one hand, such a cross-block interaction strategy eases the information flow across Transformer blocks, thus leading to more stable and effective network optimization. On the other hand, the self-cross attention map not only models intra-relations of an image, but also reflects how the learned representations evolve in the previous block. As shown in the experiments, this could facilitate representations to keep evolving in deep layers, thus improving the quality of image representation.

The key contributions of this work are as follows.

- To the best of our knowledge, the EgoTR is the first model using Transformer for cross-view geo-localization. The globally context-aware nature of the EgoTR effectively reduces visual ambiguities in cross-view geo-localization, while the positional encoding endows the EgoTR with the notion of geometry, thus decreasing ambiguities caused by geometry misalignments. Since the position embeddings are learned without imposing strong assumption on the position knowledge, the EgoTR has wider practical applicability compared with state-of-the-art models.
- We propose a novel self-cross attention mechanism, which interacts within cross-layer patches to ensure effective information flow across Transformer blocks and to encourage representations to keep evolving. This simple yet effective design consistently enhances the representation and the generalization ability of the EgoTR, without adding additional computational cost.
- Extensive experiments demonstrate that our EgoTR brings consistent and significant performance improvements for a wide range of cross-view matching tasks, including standard, fine-grained and cross-dataset cross-view geo-localization. On all these tasks, the EgoTR exhibits its superiority of learning visually discriminative and position-aware representations and achieves a new state-of-the-art performance.

2 Related Work

The key to cross-view geo-localization is to understand and correspond both image content (appearance and semantics) and spatial layout across views. To this end, existing cross-view geo-localization methods can be roughly grouped to two categories: content-based and geometry-based.

Content-based methods focus on learning image representations that are discriminative enough to distinguish between similar looking images. Leveraging on the success of CNNs, Workman and Jacobs [23] first introduce CNNs to the cross-view matching task. Later on, Hu *et al.* [8] incorporate a two-branch VGG [19] backbone network with NetVlad layers [1] to learn viewpoint-invariant representations. They also devise a weighted soft-margin triplet loss, which can speed up the network training. Sun *et al.* [20] apply the powerful ResNet [7] as backbone networks. Coupled with capsule layers [15], their proposed GeoCapsNet is capable of modeling high-level semantics. To steer where to focus in images, the attention mechanism is introduced to the field of cross-view geo-localization. Cai *et al.* [2] introduce a lightweight attention module that combines spatial and channel attention mechanisms to emphasize visually salient features. They also propose a novel reweighting loss that adaptively allocates weights to triplets according to their difficulties, thus improving the quality of network training. SAFA [16] employs a multi-head spatial attention module to aggregate informative and diverse embedding maps. While promising, few of the above methods pay enough attention to the global dependencies of cross-view images, which hinders the discriminativeness of their embedded features. Different from existing methods, this work makes the first exploration to introduce Transformer [21] to cross-view geo-localization. We demonstrate the importance of considering global dependencies for reducing visual ambiguities.

Geometry-based methods aim to correspond geometric configurations between ground and aerial images, which helps to reduce ambiguities caused by geometry misalignments. To this end, Liu and Li [9] explicitly inject per-pixel orientation information into the network. Nevertheless, this is based on the assumption of accessibility of the groundtruth orientation, which is not always satisfied in practice. Shi *et al.* [16] employ polar transform algorithm to warp satellite images so that aerial images are geometrically aligned with ground images. However, this method is only applicable to the ideal case where the ground images are orientation-aligned panoramas. Even though the dynamic similarity module proposed in [17] overcomes this limitation, the bruteforce warping strategy of the polar transform overlooks the depth of the scene content and results in obvious appearance distortions, which hinders the performance improvement. Regmi and Shah [14] attempt to tackle this problem by synthesizing the corresponding satellite image from a ground query using conditional GANs (cGANs), but the synthesized images are always granulated and lack details. In this paper, our EgoTR explicitly encodes learnable positional embeddings into the network, without imposing strong assumption. Different from the previous works [9, 16, 17], as shown in the experiments, our EgoTR not only learns relative positional information but also considers the scene context when corresponding geometric configurations across views.

3 Method: EgoTR

In this paper, we propose a novel evolving geo-localization Transformer (EgoTR) architecture with the self-cross attention mechanism for cross-view geo-localization. The following sections detail our problem setting, objective, the EgoTR architecture and our proposed self-cross attention.

3.1 Problem Formulation and Objective

The goal of cross-view geo-localization is to localize a query ground image by matching it with a set of geo-tagged aerial images. We formulate this problem in the same way as prior works [9, 18, 16, 17].

Assume we have a training set $D = \{(g_1, a_1), ..., (g_N, a_N)\}$ containing N cross-view image pairs of ground images g and aerial images a. To simplify the problem, during training phase, let each ground image g_i corresponds to only one ground-truth aerial image a_i ($i \in \{1, 2, ..., N\}$). Given a cross-view image pair (g_i, a_i), we infer the corresponding image representations as $(\mathbf{F}_i^g, \mathbf{F}_i^a)$. Then, the weighted soft-margin triplet loss [8] L of the ith exemplar can be defined as follows:

$$L = log(1 + e^{\alpha(d(\mathbf{F}_i^g, \mathbf{F}_i^a) - d(\mathbf{F}_i^g, \mathbf{F}_j^a))}) \tag{1}$$

where $j \in \{1, 2, ..., N\}$ and $j \neq i$. α is a hyperparameter used to speed up training convergence, and $d(\cdot, \cdot)$ denotes the L_2 distance.

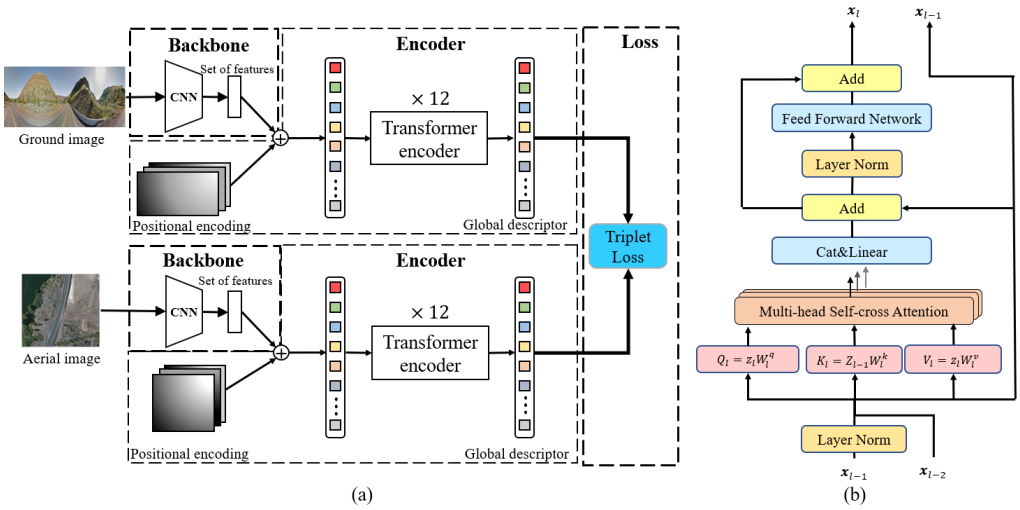


Figure 1: (a) Overview of our evolving geo-localization Transformer (EgoTR). (b) Illustration of the encoder layer with the self-cross attention in the EgoTR. $\mathbf{x}_{l-1}$ denotes the input of layer l.

3.2 Transformer for Cross-view Geo-localization

We seek to develop an EgoTR architecture that explores the global context and the positional information of cross-view images.

Preliminaries: Vision Transformer. We first describe the Vision Transformer (ViT) [4] architecture as background. Given an image, the ViT first splits it into several patches. Then, the ViT receives as input a sequence of linear projected patch embeddings $\mathbf{x} \in \mathbb{R}^{N \times D}$, where N is the number of patches, and D is the patch embedding size. After prepending a learnable class embedding $\mathbf{x}_{class} \in \mathbb{R}^{D}$, whose state at the output of the ViT is the image representation, and adding positional embeddings $\mathbf{x}_{pos}$ to $\mathbf{x}$, we gain $\mathbf{x}_0 = [\mathbf{x}_{class}; \mathbf{x}] + \mathbf{x}_{pos}$ and feed it into a L-layer Transformer encoder. Each layer consists of a Multihead Self-Attention module (MSA), Feed Forward Networks (FFN) and LayerNorm blocks (LN). Note that, the MSA is consist of multiple self-attention heads and a linear projection block. In order to make a clear comparison with our proposed self-cross attention head, we denote the input of layer l ($l \in \{1, ..., L\}$) as $\mathbf{x}_{l-1}$ and formulate a single self-attention head, the core of the vanilla MSA, as follows:

$$
\begin{aligned}
\mathbf{z}_l &= LN(\mathbf{x}_{l-1}) \\
\mathbf{Q}_l &= \mathbf{z}_l \mathbf{W}_l^q, \mathbf{K}_l = \mathbf{z}_l \mathbf{W}_l^k, \mathbf{V}_l = \mathbf{z}_l \mathbf{W}_l^v \\
\mathbf{A}_l &= softmax(\frac{\mathbf{Q}_l \mathbf{K}_l^T}{\sqrt{D}}) \mathbf{V}_l
\end{aligned}
\tag{2}
$$

where $\mathbf{W}_l^q$, $\mathbf{W}_l^k$ and $\mathbf{W}_l^v$ are linear projection matrices.

Domain-specific Transformer. The drastic domain gap between ground and aerial images hints that it is difficult to match the ground and aerial representations in the same data space. To suit the cross-view geo-localization task, we adopt a domain-specific Siamese-like architecture with two independent ViT branches of the same structure to separately learn ground and aerial image representations. The network overview is illustrated in Figure 1 (a). Each branch is a hybrid structure consisting of a ResNet backbone extracting CNN feature map from an image input and a ViT modeling global context from the CNN feature map. The linear projection of patch embedding in the ViT is applied to the CNN feature map by regarding each 1×1 feature as a patch.

Learnable positional embedding. Geometric cue can greatly simplify the cross-view geo-localization task [9, 16]. Instead of imposing a pre-defined orientation knowledge on the network, this paper applies an efficient and flexible way to endow the network with the notion of geometry. Specifically, we use learnable 1D positional embeddings in the ViT, *i.e.* $\mathbf{x}_{pos} \in \mathbb{R}^{(N+1) \times D}$. By adding the positional embeddings to the linear patch embeddings, the transformed features become position-dependent. Furthermore, since we do not impose any assumption on the position knowledge

but learn it through our learning objective, our EgoTR has wider practical applicability. As shown in the experiment, incorporating the learnable positional embeddings helps to capture relative positional information, which generalizes better to orientation-unknown images than absolute positional information. In addition, our EgoTR takes the scene content into account when corresponding cross-view geometry, which is complementary to the polar transform [16, 17] and leads to a better localization performance.

3.3 Self-cross Attention

In the vanilla ViT, the attention map is calculated independently in each layer. However, as mentioned before, such an independent learning strategy hinders the model's representation ability. To improve the quality of the learned representations, this paper proposes a novel self-cross attention mechanism to interact features between adjacent layers. Specifically, the attention map of layer l is learned not only based on $\mathbf{x}_{l-1}$, but also $\mathbf{x}_{l-2}$. Formally, in layer l, the self-cross attention can be represented as:

$$\begin{aligned} \mathbf{z}_l &= LN(\mathbf{x}_{l-1}), \mathbf{z}_{l-1} = LN(\mathbf{x}_{l-2}), \\ \mathbf{Q}_l &= \mathbf{z}_l \mathbf{W}_l^q, \mathbf{K}_l = \mathbf{z}_{l-1} \mathbf{W}_l^k, \mathbf{V}_l = \mathbf{z}_l \mathbf{W}_l^v \\ \mathbf{A}_l &= softmax(\frac{\mathbf{Q}_l \mathbf{K}_l^T}{\sqrt{D}}) \mathbf{V}_l \end{aligned} \tag{3}$$

Note that, for $l = 1$, we set $\mathbf{z}_{l-1} = LN(\mathbf{x}_{l-1})$. In Figure 1 (b), we illustrate the structure of the self-cross attention-based encoder layer.

How the self-cross attention affects feature learning. Compared to the self-attention in Eq. 2, our proposed self-cross attention creates short path between adjacent layers, thus allowing information flow effectively across layers. In fact, this shares the similar spirit with the ResNet [7]. To investigate how this affects our Transformer-based network, we plot the recall accuracy curve on CVUSA test set [24] with increasing training epochs in Figure 2. As shown, during the early training stage, the localization performance of the self-attention-based model fluctuates a lot, while our EgoTR exhibits more stable performance improvements as training processes. As a result, the stable training of the EgoTR could improves the network performance. Furthermore, by interacting cross-layer features, the attention map models how features evolve in the previous layer, which could decrease the representation similarity between layers, thus improving the network's representation ability. This is further discussed in the experiment.

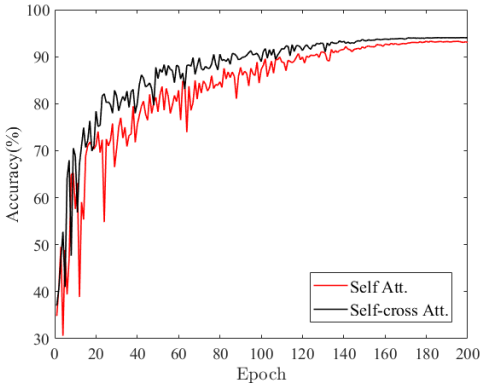


Figure 2: Recall accuracy curve on CVUSA test set with increasing training epochs.

4 Experiment

We first introduce three benchmark datasets we used to evaluate our EgoTR, evaluation protocols and implement details of our network. Then we compare our EgoTR with state-of-the-art models in Section 4.3 and present ablation studies to illustrate the advantages of the proposed EgoTR in Section 4.4. Finally, we provide qualitative results in Section 4.5 to demonstrate the effectiveness of the positional embeddings in the EgoTR.

4.1 Dataset and Evaluation Protocol

Dataset. To verify the effectiveness our model, we conduct extensive experiments on three widely used benchmarks: CVUSA [24] and CVACT [9] (including CVACT_val and CVACT_test). CVUSA dataset provides 35,532 image pairs for training and 8,884 image pairs for testing. CVACT dataset contains 35,532 pairs for training and 8,884 pairs for validation (denoted as CVACT_val). Additionally, to support fine-grained city-scale geo-localization, CVACT also provides 92,802 image pairs with accurate geo-tags for testing (denoted as CVACT_test).

Table 1: Comparisons with state-of-the-art models on CVACT_val (standard cross-view geo-localization) and CVACT_test (fine-grained geo-localization) datasets. Other results are quoted from [16] and [17]. "PT" indicates whether the model applies (w/) polar transform [16] to aerial images or not (w/o).

PT	Model	Code Length	CVACT_val				CVACT_test			
			r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
w/o	CVM-Net [8]	4096	20.15	45.00	56.87	87.57	5.41	14.79	25.63	54.53
	Liu and Li [9]	1536	46.96	68.28	75.48	92.01	19.21	35.97	43.30	60.69
	CVFT [18]	4096	61.05	81.33	86.52	95.93	26.12	45.33	53.80	71.69
	SAFA [16]	4096	78.28	91.60	93.79	98.15	-	-	-	-
	EgoTR	**768**	**83.14**	**93.84**	**95.51**	**98.40**	**58.33**	**84.23**	**88.60**	**95.83**
w/	SAFA [16]	4096	81.03	92.80	94.84	98.17	55.50	79.94	85.08	94.49
	Shi *et al.* [17]	4096	82.49	92.44	93.99	97.32	35.63	60.07	69.10	84.75
	Polar-EgoTR	**768**	**84.89**	**94.59**	**95.96**	**98.37**	**60.72**	**85.85**	**89.88**	**96.12**

Evaluation protocol. In line with [8, 9, 18, 16, 17], we evaluate our model by recall accuracy at top K (r@K for short, $K \in \{1, 5, 10, 1\%\}$), which represents the probability of correct match(es) ranking within the first K results. Note that r@1% means the recall accuracy at top 1% of test set. For CVUSA and CVACT_val, a query ground image corresponds to a single aerial image, while for CVACT_test, aerial images that locate within 5 meters of the query ground image can be seen as the correct matches.

Table 2: Comparisons with state-of-the-art methods on CVUSA [24] dataset. For all the compared methods, we cite the results from [17] and [16] if not specialized.

PT	Model	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
w/o	Workman *et al.* [23]	-	-	-	34.30
	Vo and Hays [22]	-	-	-	63.70
	Zhai *et al.* [24]	-	-	-	43.20
	CVM-Net [8]	22.47	49.98	63.18	93.62
	Liu and Li [9]	40.79	66.82	76.36	96.12
	Zheng *et al.* [25]	43.91	66.38	74.58	91.78
	Regmi and Shah [14]	48.75	-	81.27	95.98
	Siam-FCANet [2]	-	-	-	98.30
	CVFT [18]	61.43	84.69	90.49	99.02
	SAFA [16]	81.15	94.23	96.85	99.49
	EgoTR	**91.99**	**97.68**	**98.65**	**99.75**
w/	SAFA [16]	89.84	96.93	98.14	99.64
	Shi *et al.* [17]	91.93	97.50	98.54	99.67
	Polar-EgoTR	**94.05**	**98.27**	**98.99**	**99.67**

4.2 Implementation Detail

If not specified, the ground and aerial image size are set to 128×512 and 256×256, respectively. We empirically set model depth L to 12 and initialize our EgoTR with pretrained parameters on ImageNet [3]. The model is trained using AdamW [10] with cosine learning rate schedule on a 32GB NVIDIA V100 GPU. The learning rate is set to 1e-4, the weight decay is chosen to 0.03 and the batch size is 32. For the weighted soft-margin triplet loss [8], α is set to 10.

4.3 Comparing EgoTR with State-of-the-art Models

Here we compare our method with several state-of-the-art methods on CVUSA [24], CVACT_val [9] and CVACT_test [9] datasets. Unlike state-of-the-art methods that predominantly fall back on CNN, our proposed EgoTR makes the first effort to introduce Transformer to the field of cross-view geo-localization to learn globally context- and position-aware representations. Below, we verify that our EgoTR exceeds in learning visually discriminative and position-aware representations, thus achieving outstanding performance in various cross-view geo-localization tasks. Note that for fair comparison with works [16, 17] that use polar transform [16], a kind of data pre-processing algorithm, we apply the same warping strategy to aerial images before feeding them into the network (denoted as Polar-EgoTR) when comparing with these works. In this case, ground and warped aerial images are resized to 128×512.

Standard cross-view geo-localization. We first evaluate our EgoTR on standard cross-view geo-localization. Table 1 and 2 show experimental results on CVACT_val and CVUSA datasets, respectively. From the results, we could conclude that our EgoTR significantly surpasses the competing

approaches in learning visually discriminative representations and corresponding geometric configurations across views. In particular, without applying the polar transform, our EgoTR achieves r@1 of 83.14% on CVACT_val dataset compared to 78.28% obtained by the second best method, while on CVUSA dataset the EgoTR surpasses the second best method by a significant margin of 10.84 points at r@1. Moreover, when applying the polar transform, which geometrically aligns cross-view images, our EgoTR outperforms the competing methods, gaining 84.89% and 94.05% on CVACT_val and CVUSA, respectively. The results indicates that the EgoTR is capable of capturing visually discriminative features by modeling global context. Furthermore, we could also find that removing the polar transform algorithm leads to significant performance degradation in SAFA (-4.21% on CVACT_val and -8.69% on CVUSA) while less of a degradation is noted in our EgoTR (-1.75% on CVACT_val and -2.06% on CVUSA). Namely, the EgoTR does not have to rely excessively on the polar transform to establish cross-view geometric correspondence, which could save considerable image pre-processing time on large-scale datasets. This is because, adding the positional embeddings to the linear patch embeddings enables the EgoTR to learn position-aware representations and to correspond geometric configurations between ground and aerial images. Additional qualitative evidence for this is provided in Section 4.5.

Fine-grained cross-view geo-localization. To evaluate the representation ability of our model, we verify the EgoTR on fine-grained cross-view geo-localization task. Specifically, we compare the EgoTR with state-of-the-art methods on the challenging large-scale CVACT_test dataset, which is fully gps-tagged for accurate localization. Table 1 shows the experimental results. Our EgoTR performs consistently better than all the competitors, achieving 58.33% and 60.76% at r@1 without and with the polar transform, respectively. These results further demonstrate that our EgoTR has strong representation capability.

Table 3: Cross-dataset cross-view geo-localization. The results are gained by retraining and evaluating the compared models using the released codes provided by their authors.

Model	Task	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
SAFA [16]	CVUSA→CVACT	30.40	52.93	62.29	85.82
Shi *et al.* [17]		33.66	52.17	59.74	79.67
Polar-EgoTR		**47.55**	**70.58**	**77.39**	**91.39**
SAFA [16]	CVACT→CVUSA	21.45	36.55	43.79	69.83
Shi *et al.* [17]		18.47	34.46	42.28	69.01
Polar-EgoTR		**33.00**	**51.87**	**60.63**	**84.79**

Cross-dataset cross-view geo-localization. In the context of cross-view geo-localization, the transferring performance determines whether a model could be practically usable for real-life scenarios, where a query image may be dramatically different from the training ground images. As CVACT and CVUSA datasets are collected from two different countries, they have distinctly different scene styles. Based on this observation, to verify the transferring performance of our model, we train the EgoTR on CVUSA dataset and test it on CVACT_val (denoted as CVUSA→CVACT), and vice versa. Results are reported in Table 3. We could find that our EgoTR outperforms the second best model at r@1 by a large margin of 13.89 points on the CVUSA→CVACT task, while achieves 33.00% at r@1 compared to 21.45% gained by the second best model on the CVACT→CVUSA task. The transferring results demonstrate the outstanding generalization ability and practical applicability of our EgoTR.

Table 4: Ablation studies of the EgoTR.

Model	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
Polar-EgoTR	94.05	98.27	98.99	99.67
w/o self-cross att.	93.26	97.91	98.78	99.68
w/o positional emb.	90.90	97.48	98.40	99.62
EgoTR	91.99	97.68	98.65	99.75
w/o positional emb.	89.04	96.88	98.44	99.61

Comparison in terms of code length. To further illustrate the advantage of our method, we compare the EgoTR with state-of-the-art methods in terms of image descriptor dimension (also called code length) in Table 1. We observe that our EgoTR has extremely short code length of 768, which is five times shorter than that of the SAFA [16], CVFT [18] and CVM-Net [8]. A shorter code length not only implies the effective information encoding capability of the EgoTR, but also means that the

EgoTR provides an alternative with less storage space, lower computational complexity and shorter running time to cross-view geo-localization.

4.4 Ablation Study

To investigate the effectiveness of the positional embeddings and the self-cross attention mechanism, we conduct ablation studies by considering three scenarios: 1) where the positional embeddings are removed, 2) where the polar transform is removed and 3) where the self-cross attention is replaced by the self-attention.

Positional encoding. We first analyze the importance of the positional embeddings and report the ablation studies on CVUSA dataset in Table 4. From the results, we can make the following observations. First, the positional encoding endows the network with the concept of position and yields consistent improvements. In particular, adding positional embeddings improves the r@1 performance of the EgoTR from 89.04% to 91.99% and brings 3.15 points improvement at r@1 to the Polar-EgoTR. Second, we could also find that the positional embeddings are complementary to the polar transform [16, 17]. Specifically, combining the polar transform with our EgoTR improves the r@1 accuracy from 91.99% to 94.05% (+2.06%), while adding the positional embeddings to Polar-EgoTR boosts the r@1 performance from 90.90% to 94.05% (+3.15%). An interpretation of this is that, incorporating the learnable positional embeddings is capable of considering the scene content when corresponding cross-view geometric configurations, which is overlooked in polar transform algorithm as mentioned before. As a result, the polar transform and the positional encoding could jointly improve the network performance. In Section 4.5, we show qualitative results to verify this inference.

Table 5: Few-shot cross-view geo-localization on CVUSA [24]. We describe the number of training pairs of each subset and its proportion (Prop.) to the original CVUSA dataset.

Training Pairs	Prop.	Model	r@1 (%)	r@5 (%)	r@10 (%)	r@1% (%)
7,106	20%	Polar-EgoTR	**76.01**	**90.67**	**94.01**	**98.85**
		w/o self-cross att.	75.37	90.42	92.85	98.66
14,212	40%	Polar-EgoTR	**86.06**	**95.80**	**97.16**	99.38
		w/o self-cross att.	85.54	95.14	97.07	**99.42**
21,319	60%	Polar-EgoTR	**90.30**	**96.96**	**98.26**	**99.67**
		w/o self-cross att.	88.74	96.60	98.09	99.66
35,532	100%	Polar-EgoTR	**94.05**	**98.27**	**98.99**	99.67
		w/o self-cross att.	93.26	97.91	98.78	**99.68**

Self-cross attention. In Table 4, we ablate the self-cross attention mechanism by replacing it with the self-attention on CVUSA dataset. We can observe from Table 4 that, without imposing increase in model complexity, the self-cross attention mechanism improves r@1 performance from 93.26% to 94.05%, which manifests the effectiveness of the self-cross attention. Moreover, in Table 5, we ablates the self-cross attention on few-shot cross-view geo-localization task. The few-shot task aims to learn a model that can achieve generalization from only a small number of training examples [6]. To support this task, we randomly select a certain percentage (20%/40%/60%) of samples from CVUSA dataset to generate three subsets. The size of each subset and its corresponding proportion to CVUSA dataset are illustrated in Table 5. Results show that, replacing the self-cross attention with the self-attention consistently harms the network performance on few-shot task. This indicates that, the self-cross attention not only improves network performance, but also enhances its generalization ability. Furthermore, we also investigate how the self-cross attention affects feature learning. In Figure 3, we compare the final representation with the output of each intermediate layer by measuring their cosine similarity. As can be observed, replacing the self-attention with our proposed self-cross attention significantly and consistently decreases the representation similarity between layers. This result implies that the self-cross attention could prevent

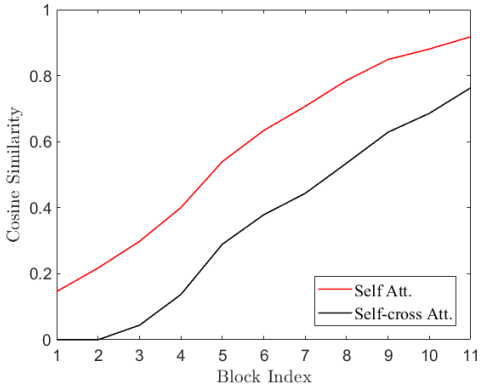
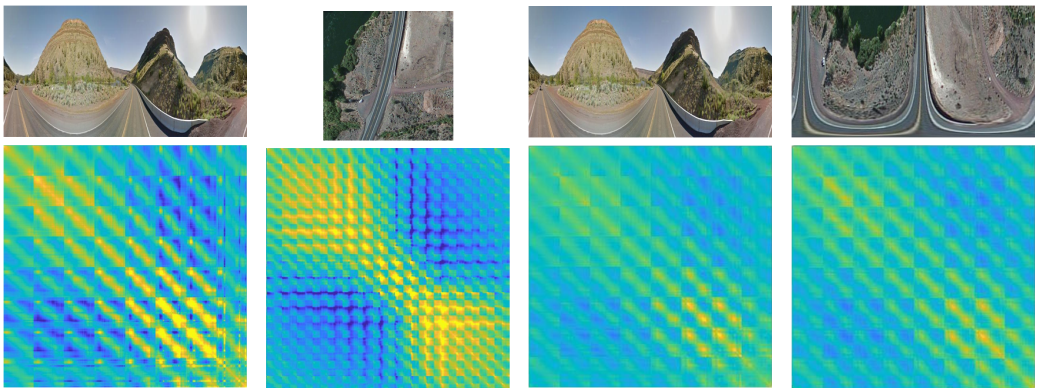


Figure 3: Cross-layer similarity between the last layer and previous layers.



(a) The ground PE of the EgoTR (b) The aerial PE of the EgoTR (c) The ground PE of the Polar-EgoTR (d) The aerial PE of the Polar-EgoTR

Figure 4: Dot products between the learnable positional embeddings (PE). Yellow indicates the two positional embeddings are closer. Better viewed in color and with zoom-in.

the learned representations of Transformer layers from being overly similar with each others and encourage the model to evolve its representations layer by layer, thus improving the performance of the final image representations.

4.5 Qualitative Analysis

We conduct detailed qualitative analysises on the learnable positional embeddings to investigate whether they encode and correspond geometric configurations across views and which positional information they can learn. To this end, we calculate dot products between two arbitrary positional embeddings of the EgoTR. From Figure 4(a) and 4(b), we could find that each positional embedding is close to its neighbors with small location offsets. This implies that, incorporating the learnable positional embeddings captures relative positional information. Additionally, we could clearly observe that, the visualization maps of the EgoTR are distinctly different across views in Figure 4(a) and 4(b), while the visualization maps of the Polar-EgoTR look similar to each other in Figure 4(c) and 4(d). Such similar results of the Polar-EgoTR are reasonable, since cross-view images are geometrically aligned by the polar transform. This result further confirms that the positional encoding could capture cross-view geometric configurations. Furthermore, in Figure 4(b), it is obvious that the positional vectors near the center of each aerial image are close to their neighbors that have fairly small distance offsets. In contrast, the positional embeddings far away from the center are close to their neighbors with relatively larger offsets. Such an offset difference is learned because of the geometric perspective of ground images. Specifically, two objects look farther away from each other in a ground image when they lie near the center of the corresponding aerial image, but look nearer to each other in the ground image when they lie far away from the center. This result unveils that the EgoTR could consider the scene content when corresponding geometric configurations across views.

5 Conclusion and Future Work

In this paper, we propose a novel EgoTR architecture capable of learning globally context- and position-aware representations. We also propose a novel self-cross attention to facilitate information flow across layers and encourage representations to keep evolving as the network layer goes deeper. Extensive experiments demonstrate that the EgoTR outperforms state-of-the-art methods in standard, fine-grained and cross-dataset cross-view geo-localization tasks. In addition, we also conduct ablation studies and qualitative analyses to verify the effectiveness of the learnable positional embeddings and the self-cross attention. One main limitation of the EgoTR is its large demand on GPU memory. Moreover, the EgoTR is built on top of the pretrained Transformer, which requires a large amount of data for training. For future work, we aim to develop a data-efficient Transformer-based model with less memory consumption for cross-view geo-localization.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8391–8400, 2019.

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[5] Christian Häne, Lionel Heng, Gim Hee Lee, Friedrich Fraundorfer, Paul Furgale, Torsten Sattler, and Marc Pollefeys. 3d visual perception for self-driving cars using a multi-camera system: Calibration, mapping, localization, and obstacle detection. *Image and Vision Computing*, 68:14–27, 2017.

[6] Jun He, Richang Hong, Xueliang Liu, Mingliang Xu, Zheng-Jun Zha, and Meng Wang. Memory-augmented relation network for few-shot learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1236–1244, 2020.

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018.

[9] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5624–5633, 2019.

[10] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. 2018.

[11] Colin McManus, Winston Churchill, Will Maddern, Alexander D Stewart, and Paul Newman. Shady dealings: Robust, long-term visual localisation using illumination invariance. In *2014 IEEE international conference on robotics and automation (ICRA)*, pages 901–906. IEEE, 2014.

[12] Sven Middelberg, Torsten Sattler, Ole Untzelmann, and Leif Kobbelt. Scalable 6-dof localization on mobile devices. In *European conference on computer vision*, pages 268–283. Springer, 2014.

[13] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9413–9422, 2020.

[14] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 470–479, 2019.

[15] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 3859–3869, 2017.

[16] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32:10090–10100, 2019.

[17] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4064–4072, 2020.

[18] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11990–11997, 2020.

[19] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[20] Bin Sun, Chen Chen, Yingying Zhu, and Jianmin Jiang. Geocapsnet: Ground to aerial view image geo-localization using capsule network. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 742–747. IEEE, 2019.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

[22] Nam N Vo and James Hays. Localizing and orienting street views using overhead imagery. In *European conference on computer vision*, pages 494–509. Springer, 2016.

[23] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015.

[24] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 867–875, 2017.

[25] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM international conference on Multimedia*, pages 1395–1403, 2020.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 1.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...

(a) Did you state the full set of assumptions of all theoretical results? [N/A]
(b) Did you include complete proofs of all theoretical results? [N/A]

3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]