

TransGeo: Transformer Is All You Need for Cross-view Image Geo-localization

Sijie Zhu, Mubarak Shah, Chen Chen

Center for Research in Computer Vision, University of Central Florida

sizhu@knights.ucf.edu, shah@crcv.ucf.edu, chen.chen@crcv.ucf.edu

Abstract

The dominant CNN-based methods for cross-view image geo-localization rely on polar transform and fail to model global correlation. We propose a pure transformer-based approach (TransGeo) to address these limitations from a different perspective. TransGeo takes full advantage of the strengths of transformer related to global information modeling and explicit position information encoding. We further leverage the flexibility of transformer input and propose an attention-guided non-uniform cropping method, so that uninformative image patches are removed with negligible drop on performance to reduce computation cost. The saved computation can be reallocated to increase resolution only for informative patches, resulting in performance improvement with no additional computation cost. This “attend and zoom-in” strategy is highly similar to human behavior when observing images. Remarkably, TransGeo achieves state-of-the-art results on both urban and rural datasets, with significantly less computation cost than CNN-based methods. It does not rely on polar transform and infers faster than CNN-based methods. Code is available at <https://github.com/Jeff-Zilence/TransGeo2022>.

1. Introduction

Image-based geo-localization aims to determine the location of a query street-view image by retrieving the most similar images in a GPS-tagged reference database. It has a great potential for noisy GPS correction [2, 33] and navigation [12, 17] in crowded cities. Due to the complete coverage and easy access of aerial images from Google Map API [1], a thread of works [10, 14, 19, 21–23, 25, 29, 35] focus on cross-view geo-localization, where the satellite/aerial images are collected as reference images for both rural [14, 34] and urban areas [29, 36]. They generally train a two-stream CNN (Convolutional Neural Network) framework employing metric learning loss [10, 35]. However, such cross-view retrieval systems suffer from the great domain gap between street and aerial views, as CNNs do not explicitly encode the position information of each view.

To bridge the domain gap, recent works apply a pre-defined polar transform [21, 22, 26] on the aerial-view images. The transformed aerial images have a similar geometric layout as the street-view query images, which results in significant boost in the retrieval performance. However, the polar transform relies on the prior knowledge of the geometry corresponding to the two views, and may fail when the street query is not spatially aligned at the center of aerial images [36] (this point is further demonstrated in Sec. 4.5).

Recently, vision transformer [7] has achieved significant performance on various vision tasks due to its powerful global modeling ability and self-attention mechanism. Although CNN-based methods are still predominant for cross-view geo-localization, we argue vision transformer is more suitable for this task due to **three advantages**: 1) Vision transformer explicitly encodes the position information, thus can directly learn the geometric correspondence between two views with the learnable position embedding. 2) The multi-head attention [28] module can model global long-range correlation between all patches starting from the first layer, while CNNs have limited receptive field [7] and only learn global information in top layers. Such strong global modeling ability can help learn the correspondence, when two objects are close in one view while far from each other in the other view. 3) Since each patch has an explicit position embedding, it is possible to apply non-uniform cropping, which removes arbitrary patches without changing the input of other patches, while CNNs can only apply uniform cropping (*i.e.* cropping a rectangle area). Such flexibility of patch selection is beneficial for geo-localization. Since some objects in aerial-view may not appear in street view due to occlusion, they can be removed with non-uniform cropping to reduce computation and GPU memory footprint, while keeping the position information of other patches.

However, vanilla vision transformer [7] (ViT) has some limitation on training data size and memory consumption, which must be addressed when applied to cross-view geo-localization. The original ViT [7] requires extremely large training datasets to achieve state-of-the-art, *e.g.* JFT-300M [7] or ImageNet-21k [5] (a super set of the origi-

nal ImageNet-1K). It does not generalize well if trained on medium-scale datasets, because it does not have inductive biases [7] inherent in CNNs, *e.g.* shift-invariance and locality. Recently, DeiT [27] applies strong data augmentation, knowledge distillation, and regularization techniques, in order to outperform CNN on ImageNet-1K [5], with similar parameters and inference throughput. However, mixup techniques used in DeiT (*e.g.* CutMix [27, 32]) are not straight-forward for metric learning losses [10].

In this paper, we propose the first pure **transformer**-based method for cross-view **geo-localization (TransGeo)**. To make our method more flexible without relying on data augmentations, we incorporate Adaptive Sharpness-Aware Minimization (ASAM) [11], which avoids overfitting to local minima by optimizing the adaptive sharpness of loss landscape and improves model generalization performance. Moreover, by analyzing the attention map of top transformer encoder, we observe that most of the occluded regions in aerial images have negligible contribution to the output. This motivates us to introduce the attention-guided non-uniform cropping, which first attends to informative image regions based on attention map of transformer encoder, then increases the resolution only on the selected regions, resulting in an “attend and zoom-in” procedure, similar to human vision. Our method achieves state-of-the-art performance with significant less computation cost (GFLOPs) than CNN-based methods, *e.g.* SAFA [21].

We summarize our contributions as follows:

- The first *pure transformer-based* method (TransGeo) for cross-view image geo-localization, without relying on polar transform or data augmentation.
- A novel attention-guided non-uniform cropping strategy that removes a large number of uninformative patches in reference aerial images to reduce computation with negligible performance drop. The performance is further improved by reallocating the saved computation to higher image resolution of the informative regions.
- State-of-the-art performance on both urban and rural datasets with less computation cost, GPU memory consumption, and inference time than CNN-based methods.

2. Related Work

Cross-view Geo-localization Existing works for cross-view geo-localization [3, 10, 13, 14, 25, 29, 30] generally adopt a two-stream CNN framework to extract different features for two views, then learn an embedding space where images from the same GPS location are close to each other. However, they fail to model the significant appearance gap between two views, resulting in poor retrieval performance. Recent methods either leverage polar transform [21, 22, 26] or add additional generative model [19, 26] (GAN [8]) to reduce domain gap by transforming images from one view to the other. SAFA [21] designs a polar transform based

on the geometric prior knowledge of the two views, so that the transformed aerial images have similar layouts as street-view images. Toker *et al.* [26] further train a generative network on the top of polar transform, so that the generated images are more realistic for matching. However, they highly rely on the geometric correspondence of two views.

On the other hand, several works start to consider practical scenarios where the street-view and aerial-view images are not perfectly aligned in terms of orientation and spatial location. Shi *et al.* [22] propose a Dynamic Similarity Matching module to account for orientation while computing the similarity of image pairs. Zhu *et al.* [35] adopt improved metric learning techniques and leverages activation map for orientation estimation. VIGOR [36] proposes a new urban dataset assuming that the query can occur at arbitrary locations in a given area, so the street-view image is not spatially aligned at the center of aerial image. In such case, polar transform may fail to model the cross-view correspondence, due to unknown spatial shift and strong occlusion. *We show that vision transformer can tackle this challenging scenario with learnable position embedding on each input patch (Sec. 4.3).*

We notice that L2LTR [31] adopts vanilla ViT [7] on the top of ResNet [9], resulting in a hybrid CNN+transformer approach. Since it adopts CNN as feature extractor, the self-attention and position embedding are only used on the high-level CNN features, which does not fully exploit the global modeling ability and position information from the first layer. Besides, as noted in their paper, it requires significantly larger GPU memory [31] and pre-training dataset than CNN-based methods, while our approach enjoys GPU memory efficiency and uses the same pre-training dataset as CNN-based methods, *e.g.* SAFA [21]. We compare with their method in Sec. 4.3. More comparisons are also included in *supplementary materials*.

Vision Transformer Transformer [28] is originally proposed for large-scale pre-training in NLP. It is firstly introduced for vision tasks in ViT [7] as vision transformer. ViT divides each input image into $k \times k$ small patches, then considers each patch as one token along with position embedding and feeds them into multiple transformer encoders. It requires extremely large training datasets to outperform CNN counterparts with similar parameters, as it does not have inductive biases inherent in CNNs. DeiT [27] is recently proposed for data-efficient training of vision transformer. It outperforms CNN counterparts on medium-scale datasets, *i.e.* ImageNet [5], by strong data augmentation and regularization techniques. A very recent work [4] further reduces the augmentations to only inception-style augmentations [24]. However, even random crop could break the spatial alignment, and previous works on cross-view geo-localization generally do not use any augmentation. *We aim to design a generic framework for cross-view*

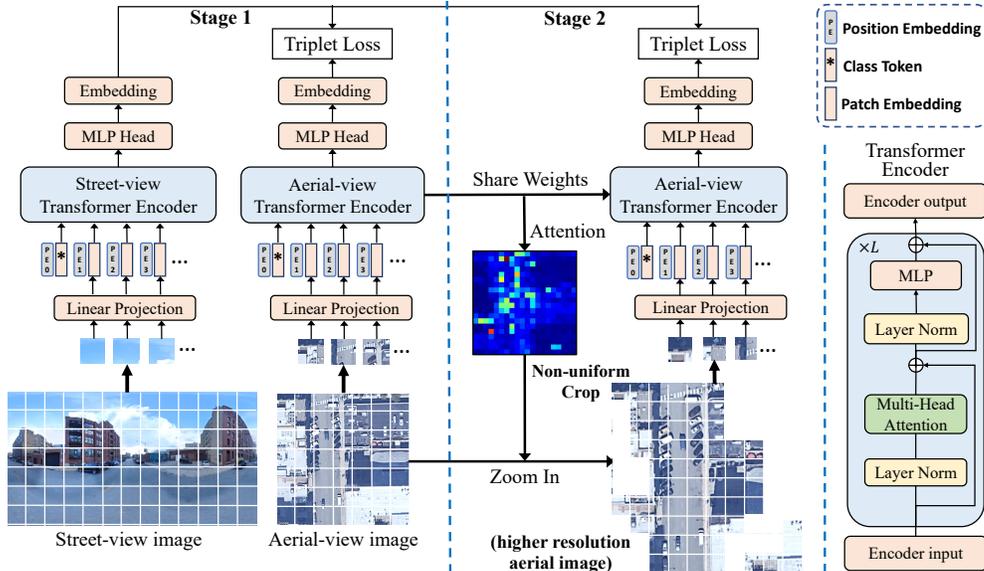


Figure 1. An overview of the proposed method. Stage-1 uses regular training by employing Eq. 1. Stage-2 follows the “attend and zoom-in” strategy by increasing the resolution of the important regions of reference aerial image, using attention-guided non-uniform cropping (Sec. 3.3). The patch size remains unchanged.

geo-localization without any augmentation, thus introduce a strong regularization technique, i.e. ASAM [11], to prevent vision transformer from overfitting.

3. Method

We first formulate the problem and present an overview of our approach in Sec. 3.1. Then in Sec. 3.2, we introduce the vision transformer components that are used in our method. We present the proposed attention-guided non-uniform cropping strategy in Sec. 3.3, which removes a large portion of patches (tokens) while maintaining retrieval performance. Finally, we introduce the regularization technique (ASAM [11]) in Sec. 3.4 for model training.

3.1. Problem Statement and Method Overview

Given a set of query street-view images $\{I_s\}$ and aerial-view reference images $\{I_a\}$, our objective is to learn an embedding space in which each street-view query I_s is close to its corresponding ground-truth aerial image I_a . Each street-view image and its ground-truth aerial image are considered as a positive pair, other pairs are considered as negative. If there are multiple aerial images covering one street-view image, e.g. VIGOR dataset [36], we consider the nearest one as the positive, and avoid sampling the other neighboring aerial images in the same batch to prevent ambiguous supervision.

Overview of Method. As shown in Fig. 1, we train two separate transformer encoders, i.e. T_s, T_a , to generate embedding features for street and aerial views, respectively.

The model is trained with soft-margin triplet loss [10]:

$$\mathcal{L}_{triplet} = \log \left(1 + e^{\alpha(d_{pos} - d_{neg})} \right). \quad (1)$$

Here d_{pos} and d_{neg} denote the squared l_2 distance of the positive and negative pairs. In a mini-batch with N street-view and aerial-view image pairs, we adopt the exhaustive strategy [20] to sample $2N(N-1)$ triplets. We apply l_2 normalization on all the output embedding features.

Fig. 1 shows the overall pipeline of our method. Stage 1 applies regular training with Eq. 1. In stage 2, we adopt the attention map of aerial image as guidance and perform non-uniform cropping (Sec. 3.3), which removes a large number of uninformative patches in reference aerial images. We then reallocate the saved computation for higher image resolution only on important regions.

3.2. Vision Transformer for Geo-localization

We briefly describe the vision transformer [7] components that are adopted in our method, i.e. patch embedding, position embedding, and multi-head attention.

Patch Embedding: Given the input images $I \in \mathbb{R}^{H \times W \times C}$, the patch embedding block converts them into a number of tokens as the input of transformer encoders. Here H, W, C denote the height, width and channel numbers of I . As shown in Fig. 1, images are first divided into N $P \times P$ patches (we use $P = 16$), $I_p \in \mathbb{R}^{N \times (P \times P \times C)}$. All the N patches are further flattened as $\mathbb{R}^{N \times P^2 C}$ and fed into the trainable linear projection layer to generate N tokens, $I_t \in \mathbb{R}^{N \times D}$. D is the feature dimension of transformer encoder.

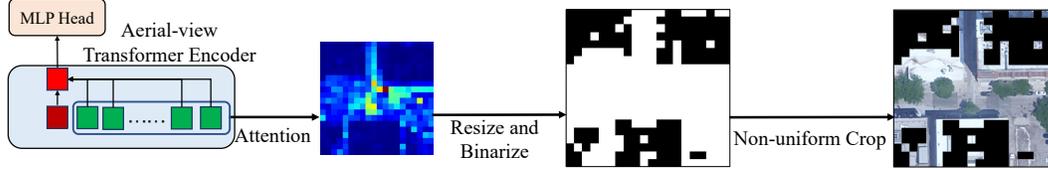


Figure 2. Pipeline of the proposed attention-guided non-uniform cropping scheme. The red box indicates the class token. The other green boxes indicate patch tokens. The patches shown in black are not selected in the input.

Class Token: In addition to the N image tokens, ViT [7] adds an additional learnable class token following BERT [6], to integrate classification information from each layer. The output class token of the last layer is then fed into an MLP (Multilayer Perceptron) head to generate the final classification vector. We use the final output vector as the embedding feature and train it with the loss in Eq. 1.

Learnable Position Embedding: Position embedding is added to each token to maintain the positional information. We adopt the learnable position embedding in ViT [7], which is a learnable matrix $\mathbb{R}^{(N+1) \times D}$ for all the $(N+1)$ tokens including class token. The learnable position embedding enables our two-stream transformer to explicitly learn the best positional encoding for each view without any prior knowledge on the geometric correspondence, thus is more generic and flexible than CNN-based methods. *The position embedding also makes it possible to remove arbitrary tokens without changing the position information of other tokens, inspiring us for employing the non-uniform cropping.*

Multi-head Attention: On the right of Fig. 1, we show the inner architecture of the transformer encoder, which is L cascaded basic transformer blocks. The key component is the multi-head attention block. It first uses three learnable linear projections to convert the input into query, key and value, denoted as Q, K, V with dimension D . The attention output is then computed as $\text{softmax}(QK^T/D)V$. A k -head attention block performs the linear projection to Q, K, V with k different heads. The attention is then performed in parallel for all the k heads. The outputs are concatenated and projected back to the model dimension D . The multi-head attention block can model strong global correlation between any two tokens starting from the first layer, which is not possible to learn in CNNs due to limited receptive field of convolution. However, note that the computation complexity is $\Theta(N^2)$, a large number of tokens will have a large computation cost. In other words, reducing the number of tokens is desirable to save computation.

3.3. Attention-guided Non-uniform Cropping

When looking for cues for image matching, humans generally take the first glance to find the most important regions, then attend to only the important regions and zoom-in to find more details with high resolution. For cross-view

geo-localization, the “attend and zoom-in” procedure can be more beneficial, because two views only share a small number of visible regions. A large number of regions in one view, e.g. roof of tall buildings in aerial view, maybe invisible in the other view, thus contribute negligibly to the final similarity as shown in Fig. 2. Those regions may be removed to reduce the computation and memory cost. However, important regions are often scattered across the image, therefore the uniform cropping (*i.e.* rectangular areas) in CNNs cannot remove scattered regions, as the cropped image must be rectangular. We thus propose the attention-guided non-uniform cropping in our pure transformer architecture.

As shown in Fig. 2, we employ the attention map in the last transformer encoder of aerial-view branch, because it represents the contribution of each token to the final output. Since only the output corresponding to class token is connected with the MLP head, we select the correlation between class token and all other patch tokens as the attention map and reshape to the original image shape. In the example of Fig. 2, the important regions mainly belong to the street area, and the other buildings occluded in street-view mostly have a low attention score. We then determine what portion of patches, β (e.g. 64%), to maintain after cropping.

To zoom-in for more detailed information, we maintain the patch size and increase the image resolution by $\sqrt{\gamma}$ times to have γ times of patches. The attention map is resized and binarized based on γ and β respectively, resulting in $\gamma\beta N$ patches after cropping (Fig. 2).

If $\beta \times \gamma = 1$, then the final number of tokens will be the same as our stage-1 baseline model. We can also use $\gamma = 1$ to merely reduce the number of tokens without increasing resolution, therefore improving the computation efficiency. In practice, the attention maps only need to be computed once and can be saved during the stage-1 training, thus do not introduce additional computation cost. Since the street-view branch is unchanged, the inference speed for street-view query is the same as the stage-1 model, which is faster than typical CNN-based methods (see details in Sec. 4.4).

3.4. Model Optimization

To train our transformer model without augmentation, we adopt a strong regularization/generalization technique, ASAM [11]. While optimizing the main loss in Eq. 1, we

also use ASAM to minimize the adaptive sharpness of the loss landscape, so that the model converges with a smooth loss curvature to achieve a strong generalization ability. For a given loss function \mathcal{L} and parameter weights $w \in \mathbb{R}^k$, the sharpness of loss is defined as:

$$\max_{|\epsilon|_2 < \rho} \mathcal{L}(w + \epsilon) - \mathcal{L}(w), \quad (2)$$

which is the maximal value in a l_2 ball region with radius ρ . ϵ is the perturbation on parameter weights w and $\|\cdot\|_2$ means l_2 norm. Kwon *et al.* [11] find that the sharpness is dependent on the scale of weights. In other words, any scaling factor A on w with no effect on loss \mathcal{L} can change the sharpness of loss. Kwon *et al.* then find a family of invertible linear operators $\{T_w \in \mathbb{R}^k | T_{Aw}^{-1} = T_w^{-1}\}$ as normalization operations to cancel out the effect of scaling A . Then the adaptive sharpness is defined as:

$$\max_{|T_w^{-1}\epsilon|_2 < \rho} \mathcal{L}(w + \epsilon) - \mathcal{L}(w). \quad (3)$$

Such scale-independent sharpness is highly beneficial for transformer, as the weight scales vary dramatically in transformer encoders due to strong self-attention with soft-max. By simultaneously minimizing the loss in Eq. 1 and adaptive sharpness in Eq. 3, we are able to overcome the over-fitting issue without using any data augmentation.

4. Experiment

4.1. Datasets and Evaluation Metrics

We conduct experiments on two city-scale datasets, *i.e.* CVUSA [34] and VIGOR [36], to evaluate our method on both rural and urban scenarios. They represent spatially aligned (CVUSA) and unaligned (VIGOR) settings as a complete coverage on popular settings and practical needs. **CVUSA:** The CVUSA (Cross-View USA) [30] dataset is originally proposed for large-scale localization across the U.S., containing more than 1 million of ground-level and aerial images. Zhai *et al.* [34] use the camera’s extrinsic parameters to align image pairs by warping the panoramas. This subset has 35,532 image pairs for training and 8,884 image pairs for testing. We use this subset in our experiments by following previous works [10, 21, 34].

VIGOR: VIGOR [36] originally contains 238,696 panoramas and 90,618 aerial images from four cities, *i.e.* Manhattan, San Francisco, Chicago, and Seattle. A balanced sampling is applied to select only two positive panoramas for each satellite image, resulting in 105,214 panoramas. VIGOR assumes that the queries can belong to arbitrary locations in the target area, thus is not spatially aligned to the center of any aerial reference images in both training and test sets. It has two evaluation protocols [36], *i.e.* same-area and cross-area. Besides, VIGOR provides the raw GPS

which allows meter-level evaluation. We follow the setting of VIGOR with both same-area and cross-area protocols.

Evaluation Metrics: We report the retrieval performance in terms of top- k recall accuracy, denoted as “R@ k ”. The k nearest reference neighbors in the embedding space are retrieved based on cosine similarity for each query. If the ground-truth reference image appears in the top k retrieved images, it is considered as correct. In addition, we compute the real-world distance between the predicted and ground-truth GPS locations as meter-level evaluation on VIGOR [36] dataset. Following VIGOR [36], we also report the hit rate, which is the percentage of top-1 retrieved reference images covering the query image (including the ground-truth).

4.2. Implementation Details

Our method is implemented in pytorch [18]. For CVUSA, panoramas and aerial images are resized to 112×616 and 256×256 before feeding into our model with batch size of 32, following [21]. For VIGOR, panoramas and aerial images respectively are resized to 640×320 and 320×320 with batch size of 16, following [36]. The patch size is 16×16 and the feature dimension is 384. We use 12 transformer encoders with 6 heads for each multi-head attention block. The model is initialized with off-the-shelf pre-trained weights [27] on ImageNet-1K [5]. We use AdamW [16] optimizer with learning rate of 0.0001 based on cosine scheduling [15]. The weight (α in Eq. 1) of soft-margin triplet loss [10] is set to 10. More details are available in [supplementary materials](#). The dimension of final embedding feature is 1,000 which is much smaller than typical CNN-based methods, *e.g.* 4,096 in SAFA [21].

4.3. Comparison with State-of-the-art

VIGOR: The proposed transformer-based method is more advantageous to VIGOR, where the two views are not perfectly aligned in terms of spatial location, due to the strong global modeling and learnable position embedding. As shown in Table 1, the proposed method significantly outperforms previous state-of-the-art methods. *The relative improvement respectively is 49.7% and 72.6% for the same-area and cross-area protocols over VIGOR on R@1*, indicating the strong learning capacity and robustness to cross-city distribution shift (cross-area setting uses different cities for training and testing).

Meter-level Evaluation: Since the final goal of localization is to get a small localization error in terms of distance (meters), we conduct the meter-level evaluation following [36]. We apply different thresholds in terms of meters and compute the corresponding accuracy when the distance between the predicted and ground-truth GPS is smaller than the threshold. As shown in Fig. 3, the proposed method significantly outperforms previous works on both settings, especially for threshold > 20 m. We get [36] a slightly

	Same-Area					Cross-Area				
	R@1	R@5	R@10	R@1%	Hit	R@1	R@5	R@10	R@1%	Hit
Siamese-VGG [35]	18.69	43.64	55.36	97.55	21.90	2.77	8.61	12.94	62.64	3.16
SAFA [21]	33.93	58.42	68.12	98.24	36.87	8.20	19.59	26.36	77.61	8.85
SAFA+Mining [36]	38.02	62.87	71.12	97.63	41.81	9.23	21.12	28.02	77.84	9.92
VIGOR [36]	41.07	65.81	74.05	98.37	44.71	11.00	23.56	30.76	80.22	11.64
Ours	61.48	87.54	91.88	99.56	73.09	18.99	38.24	46.91	88.94	21.21

Table 1. Comparison with previous works in terms of retrieval accuracy (%) on VIGOR. Hit means hit rate in [36].

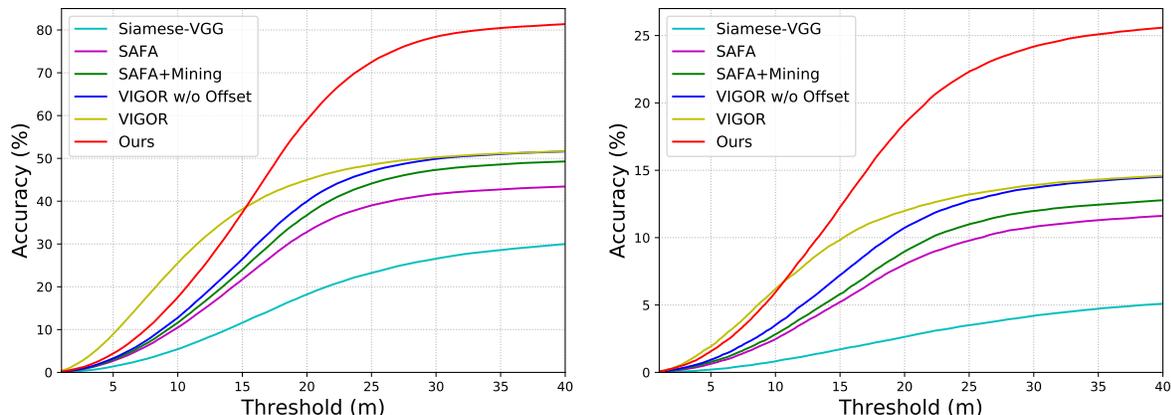


Figure 3. Same-area (left) and cross-area (right) meter-level localization accuracy of previous works and the proposed method.

Method	R@1	R@5	R@10	R@1%
CVM-Net [10]	22.47	49.98	63.18	93.62
Liu [14]	40.79	66.82	76.36	96.12
Reweight [3]	-	-	-	98.30
Regmi [19]	48.75	-	81.27	95.98
Revisit [35]	70.40	-	-	99.10
SAFA [21]	81.15	94.23	96.85	99.49
L2LTR [31]	91.99	97.68	98.65	99.75
†SAFA [21]	89.84	96.93	98.14	99.64
†Shi [22]	91.96	97.50	98.54	99.67
†Toker [26]	92.56	97.55	98.33	99.57
†L2LTR [31]	94.05	98.27	98.99	99.67
Ours	94.08	98.36	99.04	99.77

Table 2. Comparison with previous works in terms of Recall R@k (%) on CVUSA. “†” indicates methods using polar transform.

Method	GFLOPs	GPU Memory	Inference Time per Batch	R@1
†SAFA	42.24	10.82 GB	111 ms	89.84
Ours	11.32	9.85 GB	99 ms	94.08

Table 3. Comparison with SAFA [21] in terms of GFLOPs, GPU memory, inference speed and performance on CVUSA. Both methods are tested on the same GTX 1080 Ti with batch size of 32. “†” indicates methods that use polar transform.

higher accuracy on VIGOR for extremely small thresholds, due to an extra branch predicting the offset for the aerial

image. If we remove the offset from VIGOR for a fair comparison as “VIGOR w/o Offset”, then the proposed method outperforms “VIGOR w/o Offset” on all thresholds. We can also adopt the offset prediction in the future to improve the localization on small thresholds.

CVUSA: In Table 2, we compare the proposed method with previous state-of-the-art methods. Note that our method does not use polar transform, and methods with polar transform are marked with “†”. Our method achieves state-of-the-art compared to all previous works, and outperforms methods w/o polar transform by a large margin, demonstrating the superiority of pure transformer based method over CNN-based methods. Note that L2LTR [31] uses significantly larger GPU memory and pre-training dataset than the proposed method. Our method is much more efficient with better performance. Our performance can be further improved with a larger model. Detailed comparison on computation cost is provided in Sec. 4.4. *Additional results for CVACT [14], unknown orientation, limited field of view are provided in supplementary materials.*

4.4. Computational Cost

In Table 3, we provide detailed computation comparison between the proposed method and a state-of-the-art CNN-based method, *i.e.* SAFA [21]. *To the best of our knowledge, this is the first cross-view geo-localization work that reports detailed comparison of computational cost, which*

Ablation	VIGOR Same-Area				CVUSA			
	R@1	R@5	R@1%	#patches ↓	R@1	R@5	R@1%	#patches ↓
Stage-1	59.80	86.82	99.53	400	93.18	98.08	99.76	256
Stage-2 ($\beta = 0.64, \gamma = 1$)	59.44	86.32	99.50	256	93.08	97.99	99.72	163
Stage-2 ($\beta = 0.64, \gamma = 1.56$)	61.48	87.54	99.56	400	94.08	98.36	99.77	256

Table 4. Ablation study on attention-guided non-uniform cropping of our proposed method on VIGOR and CVUSA.

is an important algorithmic aspect that has been completely overlooked in the previous geo-localization literature. We select SAFA because it does not have additional blocks like [19, 26], thus has relatively low computation among all CNN-based methods. Authors in [31] report that their method requires significantly larger GPU memory and pre-training dataset (ImageNet-21K used in ViT [7]) than CNN-based methods, as it uses vanilla ViT on the top of ResNet [9]. Therefore, our method is guaranteed to be more efficient if our computation cost is less than CNN-based methods. As shown in Table 3, the computational cost (GFLOPs) of the proposed method is only 26.8% of that of SAFA [21]. It is also more efficient in terms of training GPU memory consumption, while achieving a much higher performance. In addition, the proposed method is faster than SAFA during inference, indicating its superiority for real-world applications. Since L2LTR [31] does not provide detailed computation measurement in the paper, we analyze their code and show comparison in supplementary material.

Ablation	R@1	R@5	R@1%
VIGOR Same-Area			
SAFA [21, 36]	33.93	58.42	98.24
SAFA+Polar	24.13	45.58	95.26
CVUSA			
(Ours) Stage-1	93.18	98.08	99.76
(Ours) Stage-1+Polar	93.24	98.08	99.76

Table 5. Ablation study on polar transform.

4.5. Ablation Study

Polar Transform: In Table 5, we show the effect of polar transform on both CVUSA and VIGOR. Polar transform has been shown to significantly improve CNN-based methods, but it only has marginally improvement on our pure transformer model, because the geometric information is explicitly encoded and learned in the learnable position embedding. Therefore, we do not use polar transform to maintain our pipeline simple. For VIGOR [36], the authors claim that polar transform would not work because the two views are not spatially aligned. We verify this point in Table 5, which is clear from the performance drop of SAFA when polar transform is applied. Since the center of aerial image may not be the location of street-view query, using the

center to apply polar transform can break the geometric correspondence. Therefore, we do not apply polar transform in our method for VIGOR.

Ablation	R@1	R@5	R@1%
VIGOR Same-Area			
TransGeo w/o ASAM	52.65	78.29	98.17
TransGeo	61.48	87.54	99.56
CVUSA			
TransGeo w/o ASAM	90.92	97.03	99.44
TransGeo	94.08	98.36	99.77

Table 6. Ablation study on ASAM.

ASAM: In table 6, we show the effectiveness of ASAM on both VIGOR and CVUSA. ASAM brings 8.83% and 3.16% R@1 improvement on VIGOR and CVUSA respectively. For VIGOR dataset, “TransGeo w/o ASAM” still outperforms previous methods by a large margin, which means transformer-based method has significant superiority over CNN-based method when the two views are not perfectly aligned. On CVUSA, “TransGeo w/o ASAM” performs on par with polar-transform-based methods using less computation.

Attention-guided Non-uniform Cropping: We conduct ablation study to demonstrate the effectiveness of the proposed attention-guided non-uniform cropping. As shown in Table 4, “Stage-1” does not use any cropping strategy, and is trained for the same number of epochs as the Stage-2 models. We find that simply training for more epochs (*e.g.* 200 vs 100) does not improve the performance. “Stage-2 ($\beta = 0.64, \gamma = 1$)” removes 36% of the patches (64% kept), and does not increase the resolution $\gamma = 1$. The performance only has a negligible drop, *i.e.* 0.36 for VIGOR and 0.1 for CVUSA. The results indicate that the removed patches actually are uninformative for cross-view geo-localization and the attention guidance makes sense. We then further reallocate the saved computation by increasing the resolution ($\gamma = 1.56$), resulting 1.56 times more number of patches which is the same as in the original “Stage-1” model. *With the same number of patches, we can improve the performance on both VIGOR and CVUSA.*

Learnable Position Embedding: The position embedding (abbreviated as “Pos. Emb.”) is crucial for pure

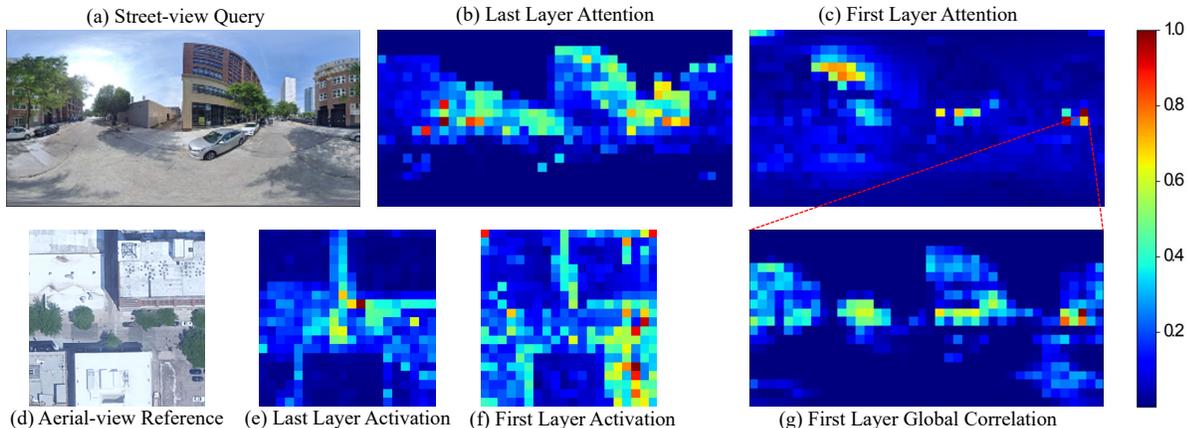


Figure 4. Visualization of the attention maps and correlation intensity in the first and last layer of our transformer encoders.

Ablation	R@1	R@5	R@1%
Fixed Pos. Emb.	42.72	68.76	94.40
Learnable Pos. Emb.	93.18	98.08	99.76

Table 7. Ablation study on different position embeddings on CVUSA in terms of Recall.

Ablation	Res.	R@1	R@5	R@1%
$\beta = 0.53, \gamma = 1$	256	92.80	97.87	99.73
$\beta = 0.64, \gamma = 1$	256	93.08	97.99	99.72
$\beta = 0.79, \gamma = 1$	256	93.10	98.03	99.75
$\beta = 0.53, \gamma = 1.88$	352	93.81	98.36	99.79
$\beta = 0.64, \gamma = 1.56$	320	94.08	98.36	99.77
$\beta = 0.79, \gamma = 1.26$	288	93.83	98.19	99.77

Table 8. Ablation study for different β and γ on CVUSA.

transformer-based methods, as there is no implicit position information (e.g. locality in CNN) for each input token. In Table 7, we compare “Learnable Pos. Emb.” with the popular predefined “Fixed Position Embedding”, i.e. Sinusoidal Embedding [28]. We use the 2D version [7] for our image-based task and all ablations are based on Stage-1 model. Results show that learnable position embedding significantly outperforms the fixed position embedding, indicating the learnable position embedding highly benefits pure transformer model when the cross-view domain gap is large.

Effect of β and γ : In Table 8, we show the effect of different β and γ , removing different number of patches and zoom-in with different resolutions (denoted as “Res.”). For each β , we use two γ values, i.e. $\gamma = 1$ and $\gamma = 1/\beta$. The results indicate that removing up to 47% of patches still yields a very small performance drop, while the higher resolution does not bring further improvement on performance, we thus select the best performed $\beta = 0.64$ as our default setting.

4.6. Visualization

In Fig. 4, we visualize the attention maps of our model on VIGOR as described in Sec. 3.3. Given a pair of street-view and aerial-view images in Fig. 4(a),(d), we generate the overall attention of each location from the first and last layers in Fig. 4(b),(c),(e),(f). The attention map of the last layer generally better aligns with the semantics of the images than the first layer and provides more high-level information that highlights informative regions. Therefore, leveraging the attention map from the last layer as guidance is reasonable. We also select the patch with maximal overall attention in Fig. 4(c) and visualize the correlation map between this patch and all patches in the first layer, as shown in Fig. 4(g). The result demonstrates that strong global correlation (i.e. high correlation scores distributed over the entire correlation map) is learned in our pure transformer model, which is a clear advantage over CNN-based methods.

5. Conclusion and Discussion

We propose the first pure transformer method (TransGeo) for cross-view image geo-localization. It achieves state-of-the-art results on both aligned and unaligned datasets, with less computational cost than CNN-based methods. The proposed method does not rely on polar transform, data augmentation, thus is generic and flexible.

One limitation of TransGeo is that it uses a two-stage pipeline. Developing one-stage generic transformer for cross-view image geo-localization would be promising for future study. Another limitation is that the patch selection simply uses the attention map which is not learnable with parameters. Better patch selection is worth exploring to focus on more informative patches. The meter-level localization could also be improved with additional offset prediction like [36] in the future.

Acknowledgement. This work is supported by the National Science Foundation under Grant No. 1910844.

References

- [1] <https://developers.google.com/maps/documentation/maps-static/intro>. 1
- [2] Eli Brosh, Matan Friedmann, Ilan Kadar, Lev Yitzhak Lavy, Elad Levi, Shmuel Rippa, Yair Lempert, Bruno Fernandez-Ruiz, Roei Herzig, and Trevor Darrell. Accurate visual localization for automotive applications. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 8391–8400, 2019. 2, 6
- [4] Xiangning Chen, Cho-Jui Hsieh, and Boqing Gong. When vision transformers outperform resnets without pre-training or strong data augmentations. *arXiv preprint arXiv:2106.01548*, 2021. 2
- [5] 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. 1, 2, 5
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4
- [7] 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. In *International Conference on Learning Representations*, 2020. 1, 2, 3, 4, 7, 8
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 2
- [9] 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. 2, 7
- [10] 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. 1, 2, 3, 5, 6
- [11] Jungmin Kwon, Jeongseop Kim, Hyunseo Park, and In Kwon Choi. Asam: Adaptive sharpness-aware minimization for scale-invariant learning of deep neural networks. *arXiv preprint arXiv:2102.11600*, 2021. 2, 3, 4, 5
- [12] Ang Li, Huiyi Hu, Piotr Mirowski, and Mehrdad Farajtabar. Cross-view policy learning for street navigation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8100–8109, 2019. 1
- [13] Tsung-Yi Lin, Serge Belongie, and James Hays. Cross-view image geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 891–898, 2013. 2
- [14] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 1, 2, 6
- [15] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 5
- [16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5
- [17] Piotr Mirowski, Matt Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Andrew Zisserman, Raia Hadsell, et al. Learning to navigate in cities without a map. In *Advances in Neural Information Processing Systems*, pages 2419–2430, 2018. 1
- [18] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32:8026–8037, 2019. 5
- [19] Krishna Regmi and Mubarak Shah. Bridging the domain gap for ground-to-aerial image matching. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 470–479, 2019. 1, 2, 6, 7
- [20] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 3
- [21] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In *Advances in Neural Information Processing Systems*, pages 10090–10100, 2019. 1, 2, 5, 6, 7
- [22] 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. 1, 2, 6
- [23] 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. 1
- [24] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2
- [25] Yicong Tian, Chen Chen, and Mubarak Shah. Cross-view image matching for geo-localization in urban environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3616, 2017. 1, 2
- [26] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the*

- IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. [1](#), [2](#), [6](#), [7](#)
- [27] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, pages 10347–10357. PMLR, 2021. [2](#), [5](#)
- [28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. [1](#), [2](#), [8](#)
- [29] 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. [1](#), [2](#)
- [30] 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. [2](#), [5](#)
- [31] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. *Advances in Neural Information Processing Systems*, 34, 2021. [2](#), [6](#), [7](#)
- [32] Sangdoon Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6023–6032, 2019. [2](#)
- [33] Amir Roshan Zamir and Mubarak Shah. Accurate image localization based on google maps street view. In *European Conference on Computer Vision*, pages 255–268. Springer, 2010. [1](#)
- [34] 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. [1](#), [5](#)
- [35] Sijie Zhu, Taojiannan Yang, and Chen Chen. Revisiting street-to-aerial view image geo-localization and orientation estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 756–765, January 2021. [1](#), [2](#), [6](#)
- [36] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#)