Cross-View Geo-Localization via Learning Disentangled Geometric Layout Correspondence

Xiaohan Zhang[1,2*], **Xingyu Li**[3*], **Waqas Sultani**[4], **Yi Zhou**[5], **Safwan Wshah**[1,2†]

[1]Department of Computer Science, University of Vermont, Burlington, USA
[2]Vermont Complex Systems Center, University of Vermont, Burlington, USA
[3] Shanghai Center for Brain Science and Brain-Inspired Technology, China
[4] Intelligent Machine Lab, Information Technology University, Pakistan
[5] NEL-BITA, School of Information Science and Technology, University of Science and Technology of China, China

Abstract

Cross-view geo-localization aims to estimate the location of a query ground image by matching it to a reference geo-tagged aerial images database. As an extremely challenging task, its difficulties root in the drastic view changes and different capturing time between two views. Despite these difficulties, recent works achieve outstanding progress on cross-view geo-localization benchmarks. However, existing methods still suffer from poor performance on the cross-area benchmarks, in which the training and testing data are captured from two different regions. We attribute this deficiency to the lack of ability to extract the spatial configuration of visual feature layouts and models' overfitting on low-level details from the training set. In this paper, we propose GeoDTR which explicitly disentangles geometric information from raw features and learns the spatial correlations among visual features from aerial and ground pairs with a novel geometric layout extractor module. This module generates a set of geometric layout descriptors, modulating the raw features and producing high-quality latent representations. In addition, we elaborate on two categories of data augmentations, (i) Layout simulation, which varies the spatial configuration while keeping the low-level details intact. (ii) Semantic augmentation, which alters the low-level details and encourages the model to capture spatial configurations. These augmentations help to improve the performance of the cross-view geo-localization models, especially on the cross-area benchmarks. Moreover, we propose a counterfactual-based learning process to benefit the geometric layout extractor in exploring spatial information. Extensive experiments show that GeoDTR not only achieves state-of-the-art results but also significantly boosts the performance on same-area and cross-area benchmarks. Our code can be found at https://gitlab.com/vail-uvm/geodtr.

Introduction

Cross-view geo-localization is defined as the estimation of the location of a ground image (also known as query image) from a set of geo-tagged aerial images (also known as reference images). The ground images are usually captured from cameras mounted on vehicles or taken by pedestrians. Cross-view geo-localization can be applied in different fields such as autonomous driving (Kim and Walter 2017), unmanned aerial vehicle navigation (Shetty and Gao 2019), and augmented reality (Chiu et al. 2018). Most of the existing cross-view geo-localization methods (Shi et al. 2019, 2020a; Yang, Lu, and Zhu 2021; Hu et al. 2018; Vo and Hays 2016; Workman, Souvenir, and Jacobs 2015; Toker et al. 2021; Shi et al. 2020b; Liu and Li 2019; Wang et al. 2021; Zheng, Wei, and Yang 2020) frame the problem as a retrieval task. These methods normally train a model to push the corresponding aerial image and ground image pairs (also known as aerial-ground pairs) closer in latent space and push the unmatched pairs further away from each other. At the deployment, the location of an aerial image with the highest similarity is the prediction for a given query ground image.

Cross-view geo-localization is considered an extremely challenging problem because of 1) the drastic change of viewpoints, 2) the difference in capturing time, 3) and the different resolutions between ground and aerial images (Wilson et al. 2021). Tackling these challenges requires a detailed understanding of the image content and the spatial configuration of visual features, e.g., buildings and roads. Most existing methods (Shi et al. 2019, 2020b; Hu et al. 2018; Lin et al. 2015; Cai et al. 2019; Vo and Hays 2016) match ground and aerial images by exploiting features extracted from convolutional neural networks (CNNs). For instance, Shi et al. (2019) directly encodes the relative positions among object features by using the Spatial-aware Position Embedding (SPE) module. These methods are limited in exploring the spatial configuration of visual features which is a global property. Recently, with the advancement of Transformer (Vaswani et al. 2017), several methods (Yang, Lu, and Zhu 2021; Zhu, Shah, and Chen 2022) explore extracting latent features from global contextual information. However, such methods solely rely on multi-head attention mechanism to implicitly explore correlations in the input features. Consequently, the correlations are unavoidably entangled in those approaches.

This paper introduces *GeoDTR* which processes the low-level feature and spatial configuration of visual features separately. To capture spatial configuration, we propose a novel geometric layout extractor sub-module. This sub-module generates a set of geometric layout descriptors that reflects the global contextual information among visual features in an image. We strengthen the quality of the geometric layout

*These authors contributed equally.
†Corresponding and senior author.

descriptors through Layout simulation, Semantic augmentation (LS) and a counterfactual (CF) training schema. LS generates different layouts for aerial and ground pairs with perturbed low-level details which improves the diversity of training aerial-ground pairs. Unlike existing data augmentation methods in cross-view geo-localization, LS *maintains the geometric/spatial correspondence during training*. Thus, it can be universally applied to any geo-localization method. Moreover, we observe that LS improves the performance on cross-area experiments because of the regularization of LS. We introduce a novel distance-based counterfactual (CF) training schema to fortify the learning of the extracted descriptors. Specifically, it provides auxiliary supervision to the geometric layout extractor to refine global contextual information. The performance of our proposed model, GeoDTR, shows a substantial increase and achieves state-of-the-art compared to other algorithms on the common cross-view geolocalization datasets, CVUSA (Workman, Souvenir, and Jacobs 2015) and CVACT (Liu and Li 2019). Our contributions can be summarized as threefold:

- We propose GeoDTR which disentangles geometric information from raw features to increase the transformer efficiency. The proposed model effectively explores the spatial configurations and low-level details, and better captures the correspondence between aerial and ground images.
- We propose layout simulation and semantic augmentation techniques that improve the performance of GeoDTR (as well as existing methods) on cross-area experiments.
- We introduce a novel counterfactual-based learning schema that guides GeoDTR to better grasp the spatial configurations and therefore produce better latent feature representations.

Related Work

Cross-view Geo-localization

Feature-based Cross-view Geo-localization Feature-based geo-localization methods extract both aerial and ground latent representations from local information using CNNs (Lin, Belongie, and Hays 2013; Lin et al. 2015; Workman, Souvenir, and Jacobs 2015). Existing works studied different aggregation strategy (Hu et al. 2018), training paradigm (Vo and Hays 2016), loss functions (i.e. HER (Cai et al. 2019)) and SEH (Guo et al. 2022) and feature transformation (i.e. feature fusion (Regmi and Shah 2019) and Cross-View Feature Transport (CVFT) (Shi et al. 2020b)). The above-mentioned feature-based methods did not fully explore the effectiveness of spatial information due to the locality of CNN which lacks of ability to explore global correlations. By leveraging the ability to capture global contextual information of the transformer, our GeoDTR learns the geometric correspondence between ground images and aerial images through a transformer-based sub-module which results in a better performance.

Geometry-based Cross-view Geo-localization Recently, learning to match the geometric correspondence between aerial and street views is becoming a hot topic. Liu and Li (2019) proposed to train the model with encoded camera orientation in aerial and ground images. Shi et al. (2019) proposed SAFA which aggregates features through its learned geometric correspondence from ground images and polar transformed aerial images. Later, the same author proposed Dynamic Similarity Matching (DSM) (Shi et al. 2020a) to geo-localizing limited field-of-view ground images by a sliding-window-like algorithm. CDE (Toker et al. 2021) combined GAN (Goodfellow et al. 2014) and SAFA (Shi et al. 2019) to learn cross-view geo-localization and ground image generation simultaneously. Despite the remarkable performance achieved by these geometric-based methods, they are limited by the nature of CNNs which explores the local correlation among pixels. On the other hand, GeoDTR not only explicitly models the local correlation but also explores the global contextual information through a transformer-based sub-module. The quality of this global contextual information is further strengthened by our CF learning schema and LS technique. Finally, benefitted from the model design, GeoDTR does not solely rely on polar transformed aerial view.

Recent researches (Yang, Lu, and Zhu 2021; Zhu, Shah, and Chen 2022) also explore to capture non-local correlations in the images. L2LTR (Yang, Lu, and Zhu 2021) studied a hybrid ViT-based (Dosovitskiy et al. 2021) methods while TransGeo (Zhu, Shah, and Chen 2022) proposed a pure transformer-based model. The above mentioned methods implicitly model the spatial information from the raw features because of solely rely on transformer. Nevertheless, our GeoDTR explicitly disentangles low-level details and spatial information from raw features. Moreover, GeoDTR has less trainable parameter than L2LTR (Yang, Lu, and Zhu 2021) and does not require the 2-stage training paradigm as proposed in TransGeo (Zhu, Shah, and Chen 2022).

Data Augmentation in Cross-view Geo-localization Data augmentation is popular in computer vision. Nonetheless, data augmentation in cross-view geo-localization is very limited due to the vulnerability of the spatial correspondence between aerial and ground images which can be easily sabotaged by a minor interference. For instance, most existing methods (Liu and Li 2019; Rodrigues and Tani 2022; Vo and Hays 2016; Cai et al. 2019) randomly rotate or shift one view while fixing the other one. On the other hand, Rodrigues and Tani (2021) randomly blackout ground objects according to their segmentation from street images. In this paper, we propose LS techniques that *maintain* geometric correspondence between images of the two views while varying geometric layout and visual features during the training phase. Our extensive experiments demonstrate that LS can significantly improve the performance on cross-area datasets not only for GeoDTR but also can be universally applied to other existing methods.

Counterfactual Learning

The idea of counterfactual in causal inference (Pearl 2009) has been successfully applied in several research areas such as explainable artificial intelligence (Byrne 2019), visual question answering (Abbasnejad et al. 2020), physics simulation (Baradel et al. 2020), and reinforcement learning (Wang

et al. 2019). In this work, we propose a novel distance-based counterfactual (CF) learning schema which strengthens the quality of learned geometric descriptors for our GeoDTR. Experiments show that the proposed CF learning schema improves the performance of GeoDTR.

Methodology

Problem Formulation

Considering a set of ground-aerial image pairs $\{(I_i^g, I_i^a)\}, i = 1, \ldots, N$, where superscripts g and a are abbreviations for ground and aerial, respectively, and N is the number of pairs. Each pair is tied to a distinct geo-location. In the cross-view geo-localization task, given a query ground image I_q^g with index q, one searches for the best-matching reference aerial image I_b^a with $b \in \{1, \ldots, N\}$.

For the sake of a feasible comparison between a ground image and an aerial image, we seek discriminative latent representations f^g and f^a for the images. These representations are expected to capture the dramatic view-change as well as the abundant low-level details, such as textual patterns. Then the image retrieval task can be made explicit as

$$b = \underset{i \in \{1,\ldots,N\}}{\arg\min}\ d(f_q^g, f_i^a), \tag{1}$$

where $d(\cdot, \cdot)$ denotes the L_2 distance. For the compactness in symbols, we will use superscript v for cases that apply to both ground (g) and aerial (a) views. We adopt this convention throughout the paper.

Geometric Layout Modulated Representations

To generate high-quality latent representations for cross-view geo-localization, we emphasize the spatial configurations of visual features as well as low-level features. The spatial configuration reflects not only the positions but also the global contextual information among visual features in an image. One could expect such geometric information to be stable during the view-change. Meanwhile, the low-level features such as color and texture, help to identify visual features across different views.

Specifically, we propose the following decomposition of the latent representation

$$f^v = \mathbf{p}^v \circ \mathbf{r}^v. \tag{2}$$

$\mathbf{p}^v = \{p_m^v\}_{m=1,\ldots,K}$ is the set of K geometric layout descriptors that summarize the spatial configuration of visual features, and $\mathbf{r}^v = \{r_j^v\}_{j=1,\ldots,C}$ denotes the raw latent representations of C channels that is generated by any backbone encoder. Both p_m^v and r_j^v are vectors in $\mathbb{R}^{H \times W}$ with H and W being the height and width of the raw latent representations, respectively. The modulation operation $\mathbf{p}^v \circ \mathbf{r}^v$ expands as

$$(\langle p_1^v, r_1^v \rangle, \ldots, \langle p_1^v, r_C^v \rangle, \ldots, \langle p_K^v, r_1^v \rangle, \ldots, \langle p_K^v, r_C^v \rangle), \tag{3}$$

where $\langle p_m^v, r_j^v \rangle$ denotes the Frobenius inner product of p_m^v and r_j^v. In this sense, the resulting $f^v \in \mathbb{R}^{CK}$ are referred to as the *geometric layout modulated representations* and will be fed to Equation (1) to retrieve the best-matching aerial images. Our model design closely follows the above decomposition.

GeoDTR Model

Model Overview GeoDTR (see Figure 1 (a)) is a siamese neural network including two branches for the ground and the aerial views, respectively. Within a branch, there are two distinct processing pathways, i.e., the backbone feature pathway and the geometric layout pathway.

In the backbone feature pathway, a CNN backbone encoder processes the input image to generate raw latent representations $\mathbf{r}^v$ where $v = g$ or $v = a$. Due to the nature of the CNN backbone, these representations carry the positional information as well as the low-level feature information.

The geometric layout pathway is devoted to exploring the global contextual information among visual features. This pathway includes a core sub-module called the geometric layout extractor, which generates a set of geometric layout descriptors $\mathbf{p}^v$ based on the raw latent representations $\mathbf{r}^v$. These descriptors will modulate $\mathbf{r}^v$, integrating the geometric layout information therein. With a stand-alone treatment of the geometric layout, one avoids introducing undesired correlations among the low-level features from different visual features. In the following, we will describe the key components of GeoDTR in detail.

Geometric Layout Extractor This sub-module mines the global contextual information among the visual features and produce effective geometric layout descriptors. Despite the change in appearance across views, the arrangement of visual features remains largely intact. Hence, integrating the geometric layout information into the latent representations f^v would improve its discriminative power for cross-view geo-localization. Note that the geometric layout is a global property in the sense that it captures the spatial configuration of multiple visual features at different positions in the ground and aerial images. For example, a single visual feature can span across the image, such as the road. As a result, the geometric layout descriptors should be able to grasp the global correlation among visual features.

In order to accomplish this goal, the geometric layout extractor is built on top of the standard transformer. As shown in Figure 1 (b), the network consists of a max-pooling layer along channels, a transformer module, and two embedding layers that locate before and after the transformer. The max-pooling layer produces a saliency map $M \in \mathbb{R}^{H \times W}$. Then, M is processed by the first embedding layer, projected into K embedding vectors $E = [e_1, e_2, ..., e_k]$ where for each embedding vector the dimension is $(H \times W)/2$. On top of the standard learnable positional embedding (PE) E_{pe} (Dosovitskiy et al. 2021), we introduce an extra index-aware positional embedding adding to E

$$\hat{E} = E + \text{HardTanh}(E_{pe} + W_{LN} M^{idx}), \tag{4}$$

where $M_{mj}^{idx} = \frac{1}{C} \arg\max_{c \in 1,\ldots,C} r_{mj}^c$. W_{LN} is a learnable linear transform that maps M^{idx} to K distinct subspaces, $W_{LN} \in \mathbb{R}^{(H \times W) \times (K \times \frac{H \times W}{2})}$. The transformer explores correlations among $\hat{E}$. After projecting by the second embedding layer, our geometric layout extractor generates a set of K geometric layout descriptors $\mathbf{p}$. The detailed model settings can be found in the supplementary material.

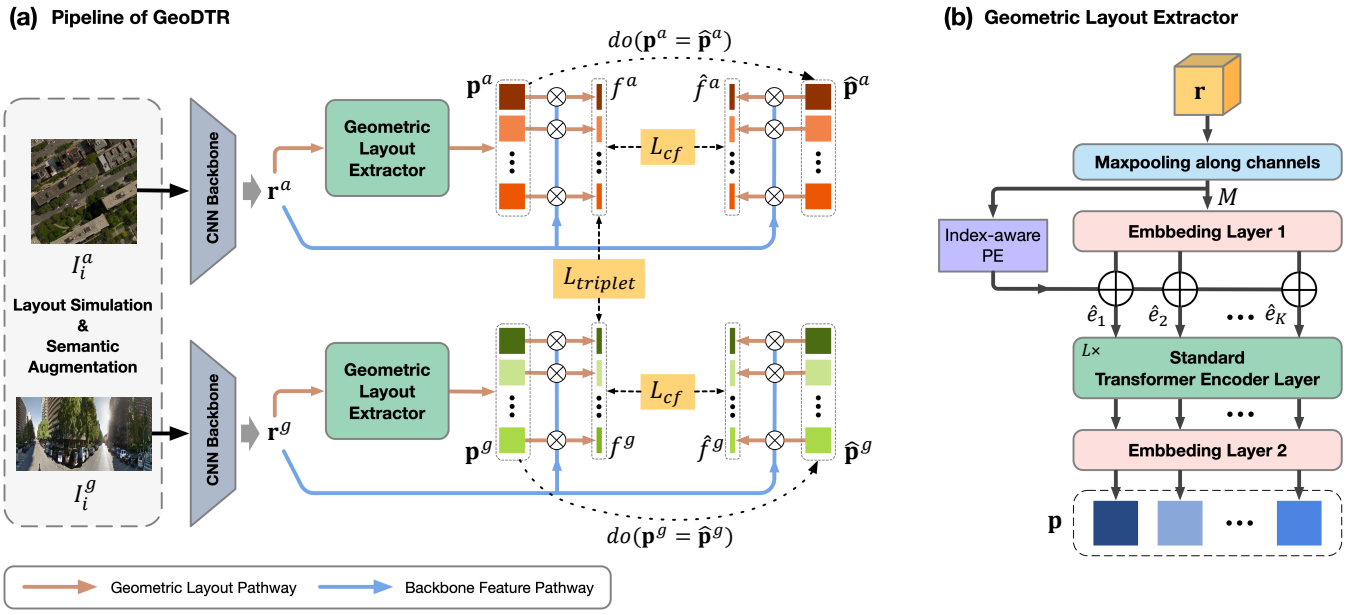

Figure 1: (a) The overview pipeline of our proposed model GeoDTR. (b) Illustration of our proposed geometric layout extractor.

Counterfactual-based Learning Schema Due to the absence of ground truth geometric layout descriptors, the submodule $\mathcal{G}^v(\cdot)$ would only receive indirect and insufficient supervision during training. Inspired by (Rao et al. 2021), we propose a counterfactual-based (CF-based) learning process. Specifically, we apply an intervention $do(\mathbf{p}^v = \hat{\mathbf{p}}^v)$ which substitutes $\mathbf{p}^v$ for a set of imaginary layout descriptors $\hat{\mathbf{p}}^v$ in Equation (2). This results in an imaginary representation $\hat{f}^v$. Elements of $\hat{\mathbf{p}}^v$ are drawn from the uniform distribution $U[-1, 1]$. In order to penalize $\hat{\mathbf{p}}^v$ and encourage $\mathbf{p}^v$ to capture more distinctive geometric clues, we maximize the distance between f^v and $\hat{f}^v$ by minimizing our proposed counterfactual loss

$$L_{cf}^v = \log\left(1 + e^{-\beta^v[d(f^v, \hat{f}^v)]}\right), \tag{5}$$

where β^v is a parameter to tune the convergence rate. The counterfactual loss provides a weakly supervision signal to the layout descriptors p via penalizing the imaginary descriptors $\hat{p}$. In this way, the model can be away from apparently "wrong" solution and learn a better latent feature representation. Besides the counterfactual loss, we also adopt the weighted soft margin triplet loss which pushes the matched pairs closer and unmatched pairs further away from each other

$$L_{triplet} = \log\left(1 + e^{\alpha[d(f_m^g, f_m^a) - d(f_m^g, f_n^a)]}\right), \tag{6}$$

where α is a hyperparameter that controls the convergence of training. $m, n \in \{1, 2, \ldots, N\}$ and $m \neq n$. Our final loss is

$$L = L_{triplet} + L_{cf}^a + L_{cf}^g. \tag{7}$$

Layout Simulation and Semantic Augmentation

In this paper, we elaborately design two categories of augmentations, i.e., Layout simulation and Semantic augmentation

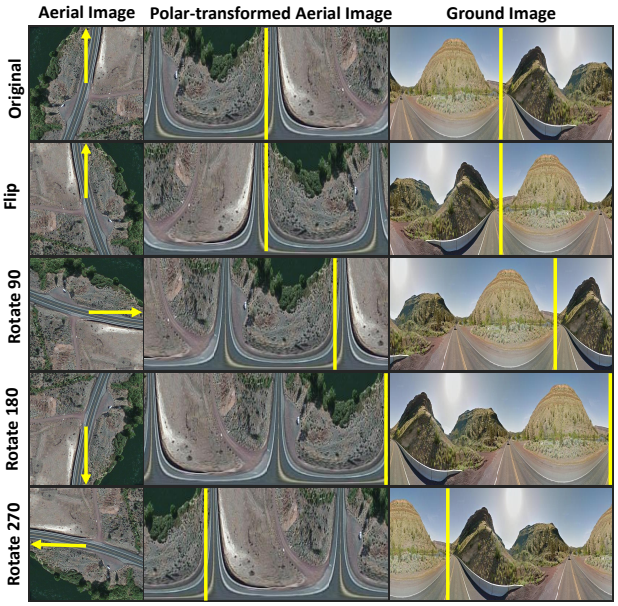

Figure 2: Illustration of the layout simulation. From left to right are aerial image, polar transformed aerial image, and ground image. The yellow arrows and lines indicate the north direction.

(LS) to help improve the quality of extracted layout descriptors and the generalization of cross-view geo-localization models.

Layout Simulation It is a combination of a random flip and a random rotation ($90°, 180°,$ or $270°$) that *synchronously* applies to ground truth aerial and ground images. In this

manner, low-level details are maintained, but the geometric layout is modified. As illustrated in Figure 2, layout simulation can produce matched aerial-ground pairs with a different geometric layout.

Semantic Augmentation Semantic augmentation randomly modifies the low-level features in aerial and ground images *separately*. We employ color jitter to modify the brightness, contrast, and saturation in images. Moreover, we also randomly apply Gaussian blur and randomly transform images to grayscale images or posterized images.

Note that unlike previous data augmentation methods, our LS does not break the geometric correspondence among visual features in the two views. Our experiments show that LS greatly improves GeoDTR on cross-area performance while hardly weakening the same-area performance. Applying LS to the existing methods also improves their performance in cross-area experiments. For more details on this, refer to the supplementary material.

Experimental Results

Method	R@1	R@5	R@10	R@1%
FusionGAN	48.75%	-	81.27%	95.98%
CVFT	61.43%	84.69%	90.49%	99.02%
SAFA	81.15%	94.23%	96.85%	99.49%
SAFA†	89.84%	96.93%	98.14%	99.64%
DSM†	91.93%	97.50%	98.54%	99.67%
CDE†	92.56%	97.55%	98.33%	99.57%
L2LTR	91.99%	97.68%	98.65%	99.75%
L2LTR†	94.05%	98.27%	98.99%	99.67%
TransGeo	94.08%	98.36%	99.04%	99.77%
SEH†	<u>95.11%</u>	98.45%	99.00%	99.78%
Ours w/ LS	93.76%	<u>98.47%</u>	<u>99.22%</u>	<u>99.85%</u>
Ours w/ LS†	**95.43%**	**98.86%**	**99.34%**	**99.86%**

Table 1: Comparison between the proposed GeoDTR and baseline methods on CVUSA dataset. † represents that polar transformation is applied to aerial images. The best results are in bold. The second-best results are underlined.

Experiment Settings

Dataset To evaluate the effectiveness of GeoDTR, we conduct extensive experiments on two datasets, CVUSA (Workman, Souvenir, and Jacobs 2015), and CVACT (Liu and Li 2019). Both CVUSA and CVACT contain $35,532$ training pairs. CVUSA provides $8,884$ pairs for testing and CVACT has the same number of pairs in its validation set (CVACT_val). Besides, CVACT provides a challenging and large-scale testing set (CVUSA_test) which contains $92,802$ pairs. In CVUSA, we identify 762 and 43 repeated pairs in the original training set and testing set, respectively. We remove the repeated pairs in the training set but keep the testing set unchanged for fair comparisons. Please refer to the supplementary material for more information of the duplicate pairs in CVUSA dataset.

Evaluation Metric Similar to existing methods (Shi et al. 2019; Toker et al. 2021; Yang, Lu, and Zhu 2021; Hu et al. 2018; Liu and Li 2019; Shi et al. 2020a), we choose to use recall accuracy at top K (R@K) for evaluation purpose. R@K measures the probability of the ground truth aerial image ranking within the first K predictions given a query image. In the following experiments, we evaluate the performance of all methods on R@1, R@5, R@10, and R@1%.

Implementation Detail We employ a ResNet-34 (He et al. 2016) as the backbone for a fair comparison with other baselines. α and β are set to 10 and 5 respectively. We train the model on a single Nvidia V100 GPU for 200 epochs with AdamW (Loshchilov and Hutter 2017) optimizer. For more information (i.e. LS techniques and latent feature dimensions, etc.), please refer to the supplementary material.

Same-area Experiment

We first evaluate GeoDTR on the same-area cross-view geo-localization tasks in which training and testing data are captured from the same region. The results on CVUSA (Workman, Souvenir, and Jacobs 2015) and CVACT (Liu and Li 2019) benchmarks are shown in Table 1 and Table 2, respectively. For a fair and complete comparison, we present the performance of GeoDTR trained with and without polar transformation (PT) on aerial images. Specifically, in the CVUSA experiments (Table 1), with PT, GeoDTR achieves the state-of-the-art (SOTA) result. Without PT, our GeoDTR exceeds TransGeo (Zhu, Shah, and Chen 2022) on R@5, R@10, and R@1% and achieve comparable results on R@1.

As shown in Table 2, GeoDTR also achieves substantial improvement in performance on CVACT. To be noticed, GeoDTR achieves 64.52% on R@1 of CVACT_test which is a *3.23%* increase from previous SOTA (CDE (Toker et al. 2021)) on this highly challenging benchmark. Furthermore, we also observe that when training without polar transformation, GeoDTR only suffers a minor decrease in performance (1.67% on R@1 of CVUSA, 0.78% on R@1 of CVACT_val, and 1.56% on R@1 of CVACT_test). We attribute this to the geometric layout descriptors that can adapt to the non-polar-transformed aerial inputs and capture the spatial configuration. More qualitative analyses on geometric layout descriptors are discussed in the later sections. The results of same-area experiments demonstrate the superiority of GeoDTR.

Cross-area Experiment

To further evaluate the generalization of GeoDTR on unseen scenes, we conduct the cross-area experiments, i.e., training on CVUSA while testing on CVACT (CVUSA $\rightarrow$ CVACT) and vice versa (CVACT $\rightarrow$ CVUSA). The results are summarized in Table 3. On CVUSA $\rightarrow$ CVACT, GeoDTR achieves 53.16% on R@1 which significantly exceeds the current SOTA (Yang, Lu, and Zhu 2021). Since CVACT is densely sampled from a single city, its images might share more common visual features. Hence, we consider CVACT $\rightarrow$ CVUSA to be a more challenging task. We observe that GeoDTR outperforms all other methods by a substantial amount. From the analyses in later sections, this significant improvement in cross-area performance comes from the cooperation between

Method	CVACT_val				CVACT_test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
CVFT	61.05%	81.33%	86.52%	95.93%	26.12%	45.33%	53.80%	71.69%
SAFA	78.28%	91.60%	93.79%	98.15%	-	-	-	-
SAFA†	81.03%	92.80%	94.84%	98.17%	55.50%	79.94%	85.08%	94.49%
DSM†	82.49%	92.44%	93.99%	97.32%	35.63%	60.07%	69.10%	84.75%
CDE†	83.28%	93.57%	95.42%	98.22%	61.29%	85.13%	89.14%	98.32%
L2LTR	83.14%	93.84%	95.51%	98.40%	58.33%	84.23%	88.60%	95.83%
L2LTR†	84.89%	94.59%	95.96%	98.37%	60.72%	85.85%	89.88%	96.12%
TransGeo	84.95%	94.14%	95.78%	98.37%	-	-	-	-
SEH†	84.75%	93.97%	95.46%	98.11%	-	-	-	-
Ours w/ LS	85.43%	94.81%	96.11%	98.26%	62.96%	87.35%	90.70%	98.61%
Ours w/ LS†	**86.21%**	**95.44%**	**96.72%**	**98.77%**	**64.52%**	**88.59%**	**91.96%**	**98.74%**

Table 2: Comparison between our GeoDTR w/ LS and baseline methods on CVACT dataset. Notations are the same as Table 1.

Model	Task	R@1	R@5	R@10	R@1%
SAFA†	CVUSA ↓ CVACT	30.40%	52.93%	62.29%	85.82%
DSM†		33.66%	52.17%	59.74%	79.67%
L2LTR†		47.55%	70.58%	77.39%	91.39%
TransGeo		37.81%	61.57%	69.86%	89.14%
Ours w/ LS		43.72%	66.99%	74.61%	91.83%
Ours w/ LS†		**53.16%**	**75.62%**	**81.90%**	**93.80%**
SAFA‡	CVACT ↓ CVUSA	21.45%	36.55%	43.79%	69.83%
DSM†		18.47%	34.46%	42.28%	69.01%
L2LTR†		33.00%	51.87%	60.63%	84.79%
TransGeo		18.99%	38.24%	46.91%	88.94%
Ours w/ LS		29.85%	49.25%	57.11%	82.47%
Ours w/ LS†		**44.07%**	**64.66%**	**72.08%**	**90.09%**

Table 3: Comparison between GeoDTR w/ LS and baselines on cross-area benchmarks. Notations are the same as Table 1.

	Same-area			Cross-area		
	R@1	R@5	R@1%	R@1	R@5	R@1%
L2LTR	94.05%	98.27%	99.67%	47.55%	70.58%	91.39%
L2LTR‡	93.62%	98.46%	99.77%	52.58%	**75.81%**	93.51%
Ours	95.23%	98.71%	99.79%	47.79%	70.52%	92.20%
Ours‡	**95.43%**	**98.86%**	**99.86%**	**53.16%**	75.62%	**93.80%**

(a) Models are trained on CVUSA dataset.

	Same-area			Cross-area		
	R@1	R@5	R@1%	R@1	R@5	R@1%
L2LTR	84.89%	94.59%	98.37%	33.00%	51.87%	84.79%
L2LTR‡	83.49%	94.93%	98.68%	37.69%	57.78%	89.63%
Ours	**87.42%**	95.37%	98.65%	29.13%	47.86%	81.09%
Ours‡	86.21%	**95.44%**	**98.77%**	**44.07%**	**64.66%**	**90.09%**

(b) Models are trained on CVACT dataset.

Table 4: Comparison between L2LTR (Yang, Lu, and Zhu 2021) and GeoDTR with or without the proposed LS technique. Results are shown for both cases when the models are trained on CVUSA and CVACT datasets. ‡ represents that LS is applied during training. The best results are in bold.

the geometric layout descriptors and our LS technique. The geometric layout descriptors efficiently grasp the spatial correlation among visual features and the LS technique helps to alleviate the model's overfitting to low-level details.

Qualitative Study of Geometric Descriptors

As discussed in our description of geometric layout extractor, our model learns to capture the geometric layout and then produces modulated latent representations upon that. The geometric layout could be considered as a fingerprint of the ground (aerial) image and should remain mostly the same under view-change. Consequently, for a ground-aerial pair, one would expect to see similar patterns in their geometric layout descriptors. To justify this point and fully demonstrate the power of our model, we visualize the ground and aerial descriptors in Figure 3 for cases when GeoDTR is trained with polar transformed aerial images and normal aerial images (without polar transformation), respectively.

In Figure 3(a), we first note a strong alignment between descriptors of a given ground-aerial pair. To better visualize, we also present the difference between the corresponding descriptors in the third column of the figure. It is clear to see that the corresponding descriptors possess very similar values apart from the ones located at a narrow strip near the top of each pair of descriptors.

More strikingly, such an alignment still exists when our model is trained with normal ground images. In Figure 3(b), we unroll the descriptors for the normal aerial image by polar transformation. We observe that apart from the deformation brought by the polar transformation, the locations of salient patterns in the aerial descriptors match those in the corresponding ground descriptors. This indicates the ability of GeoDTR to grasp the geometric correspondence even without the guidance of polar transformation.

Ablation Study

LS Technique It is worth emphasizing that our LS stands as a generic technique to improve the generalization of cross-view geo-localization models. To illustrate this point, we apply the LS technique to L2LTR (Yang, Lu, and Zhu 2021) by comparing the recall accuracy training with or without LS.

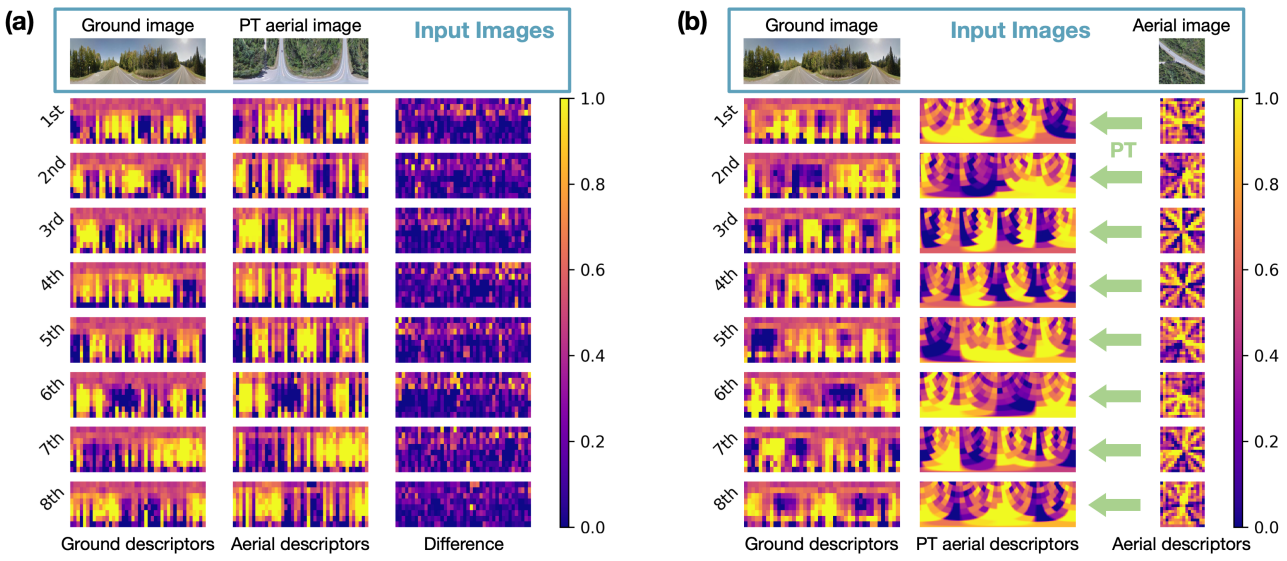

Figure 3: Visualization of learned descriptors from GeoDTR trained *with* (a) and *without* (b) polar transformation. Notice the *alignment* between ground descriptors and aerial / PT aerial descriptors.

	Configuration	R@1	R@5	R@10	R@1%
Same-area	(1, CF, w/ LS)	93.44%	98.18%	99.05%	99.80%
	(4, CF, w/ LS)	94.80%	98.64%	99.30%	99.82%
	(8, CF, w/ LS)	95.43%	98.86%	99.34%	99.86%
	(8, noCF, w/ LS)	95.06%	98.72%	99.30%	99.85%
	(8, noCF, w/o LS)	94.83%	98.59%	99.28%	99.80%
	(8, CF, w/o LS)	95.23%	98.71%	99.26%	99.79%
Cross-area	(1, CF, w/ LS)	39.93%	63.54%	71.59%	88.92%
	(4, CF, w/ LS)	46.36%	69.37%	76.46%	91.27%
	(8, CF, w/ LS)	53.16%	75.62%	81.90%	93.80%
	(8, noCF, w/ LS)	49.18%	71.96%	79.28%	92.84%
	(8, noCF, w/o LS)	43.71%	66.87%	74.68%	90.66%
	(8, CF, w/o LS)	47.79%	70.52%	77.52%	92.20%

Table 5: Performance of GeoDTR under different configurations, including the number of descriptors K, with or without counterfactual loss (CF or noCF), and with or without LS (w/ LS or w/o LS). Results are shown for both cases when our model is trianed on CVUSA and CVACT datasets.

The comparison is presented in Table 4, which also includes the same ablation on our GeoDTR. We notice that LS greatly boosts the performance of L2LTR on the cross-area benchmark with only a minor decrease on the same-area benchmark. This indicates the effectiveness of LS to improve cross-view geo-localization models in capturing clues for geo-localizing in unseen scenes. Note that even with the benefits from LS, GeoDTR w/ LS still *outperforms* L2LTR w/ LS on the overall performance in both same-area and cross-area benchmarks. The complete comparison with other existing models (i.e. SAFA) can be found in the supplementary material.

Geometric Layout Descriptors To demonstrate the benefit of geometric layout descriptors to GeoDTR, we conduct ablation experiments with the different number of descriptors K. The results are included in the first three lines of the upper and bottom parts of the ablation study (Table 5). We observe that with more descriptors, the performance is *constantly* improved, especially the cross-area ones. This implies the importance of the geometric layout for the cross-area task. Moreover, there is a notable gap in the cross-area performance between $K = 1$ and $K = 8$ cases. This gap highlights the substantial contribution of the geometric layout descriptors *in addition* to the LS technique, and, thus, reflects the effectiveness of our model design.

CF Learning Schema The effects of the CF-based learning process are shown in the last four lines in the upper and bottom parts of Table 5. We find that CF-based learning boosts the recall accuracy except for a few limited cases. The improvement is more evident in the cross-area performance when our model is trained on the CVUSA dataset. To be noticed, the value of R@1 and R@5 increase from 49.18% to 53.16% and from 71.96% to 75.62%, respectively.

Conclusion and Future Works

To address the challenges in cross-view geo-localization, we propose GeoDTR which disentangles geometric layout from raw input features and better explores the spatial correlations among visual features. In addition, we introduce layout simulation and semantic augmentation which improve the generalization of GeoDTR and other existing cross-view geo-localization models. Moreover, a novel counterfactual-based learning process is introduced to train GeoDTR. Extensive experiments demonstrate the superiority of GeoDTR on standard, fine-grained, and cross-area cross-view geo-localization tasks. Presently, the interpretation of geometric layout descriptors in our model has not been fully explored. In the future, we will keep investigating their properties and work towards more explainable models.

Acknowledgements

This project has been supported in part by VTrans and NOAA Award No. NA22NWS4320003 (subaward no. A22-0303-S001). Computations were performed on the Vermont Advanced Computing Core, supported in part by NSF award No.OAC-1827314. 04.
X. Y. L. and Y. Z. were supported by grants from the National Science and Technology Innovation 2030 Project of China (2021ZD0202600) and the National Science Foundation of China (U22B2063).

References

Abbasnejad, E.; Teney, D.; Parvaneh, A.; Shi, J.; and Hengel, A. v. d. 2020. Counterfactual Vision and Language Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Baradel, F.; Neverova, N.; Mille, J.; Mori, G.; and Wolf, C. 2020. CoPhy: Counterfactual Learning of Physical Dynamics. In *International Conference on Learning Representations*.

Byrne, R. M. 2019. Counterfactuals in Explainable Artificial Intelligence (XAI): Evidence from Human Reasoning. In *IJCAI*, 6276–6282.

Cai, S.; Guo, Y.; Khan, S.; Hu, J.; and Wen, G. 2019. Ground-to-Aerial Image Geo-Localization With a Hard Exemplar Reweighting Triplet Loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chiu, H.-P.; Murali, V.; Villamil, R.; Kessler, G. D.; Samarasekera, S.; and Kumar, R. 2018. Augmented Reality Driving Using Semantic Geo-Registration. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, 423–430.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *Advances in neural information processing systems*, 27.

Guo, Y.; Choi, M.; Li, K.; Boussaid, F.; and Bennamoun, M. 2022. Soft Exemplar Highlighting for Cross-View Image-Based Geo-Localization. *IEEE transactions on image processing*, 31: 2094–2105.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Hu, S.; Feng, M.; Nguyen, R. M.; and Lee, G. H. 2018. 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*, 7258–7267.

Kim, D.-K.; and Walter, M. R. 2017. Satellite image-based localization via learned embeddings. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2073–2080.

Lin, T.-Y.; Belongie, S.; and Hays, J. 2013. Cross-View Image Geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lin, T.-Y.; Cui, Y.; Belongie, S.; and Hays, J. 2015. Learning Deep Representations for Ground-to-Aerial Geolocalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, L.; and Li, H. 2019. Lending Orientation to Neural Networks for Cross-View Geo-Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Pearl, J. 2009. *Causality: Models, Reasoning and Inference*. USA: Cambridge University Press, 2nd edition. ISBN 052189560X.

Rao, Y.; Chen, G.; Lu, J.; and Zhou, J. 2021. Counterfactual Attention Learning for Fine-Grained Visual Categorization and Re-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1025–1034.

Regmi, K.; and Shah, M. 2019. Bridging the Domain Gap for Ground-to-Aerial Image Matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Rodrigues, R.; and Tani, M. 2021. Are These From the Same Place? Seeing the Unseen in Cross-View Image Geo-Localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3753–3761.

Rodrigues, R.; and Tani, M. 2022. Global Assists Local: Effective Aerial Representations for Field of View Constrained Image Geo-Localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 3871–3879.

Shetty, A.; and Gao, G. X. 2019. UAV Pose Estimation using Cross-view Geolocalization with Satellite Imagery. In *2019 International Conference on Robotics and Automation (ICRA)*, 1827–1833.

Shi, Y.; Liu, L.; Yu, X.; and Li, H. 2019. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32: 10090–10100.

Shi, Y.; Yu, X.; Campbell, D.; and Li, H. 2020a. 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 (CVPR)*.

Shi, Y.; Yu, X.; Liu, L.; Zhang, T.; and Li, H. 2020b. Optimal Feature Transport for Cross-View Image Geo-Localization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07): 11990–11997.

Toker, A.; Zhou, Q.; Maximov, M.; and Leal-Taixe, L. 2021. 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 (CVPR)*, 6488–6497.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vo, N. N.; and Hays, J. 2016. Localizing and Orienting Street Views Using Overhead Imagery. In Leibe, B.; Matas, J.; Sebe, N.; and Welling, M., eds., *Computer Vision – ECCV 2016*, 494–509. Cham: Springer International Publishing. ISBN 978-3-319-46448-0.

Wang, T.; Zheng, Z.; Yan, C.; Zhang, J.; Sun, Y.; Zheng, B.; and Yang, Y. 2021. Each Part Matters: Local Patterns Facilitate Cross-view Geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 1–1.

Wang, Y.; Wan, Y.; Zhang, C.; Bai, L.; Cui, L.; and Yu, P. 2019. Competitive multi-agent deep reinforcement learning with counterfactual thinking. In *2019 IEEE International Conference on Data Mining (ICDM)*, 1366–1371. IEEE.

Wilson, D.; Zhang, X.; Sultani, W.; and Wshah, S. 2021. Visual and Object Geo-localization: A Comprehensive Survey. *arXiv preprint arXiv:2112.15202*.

Workman, S.; Souvenir, R.; and Jacobs, N. 2015. Wide-Area Image Geolocalization With Aerial Reference Imagery. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Yang, H.; Lu, X.; and Zhu, Y. 2021. Cross-view Geo-localization with Layer-to-Layer Transformer. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 29009–29020. Curran Associates, Inc.

Zheng, Z.; Wei, Y.; and Yang, Y. 2020. University-1652: A Multi-View Multi-Source Benchmark for Drone-Based Geo-Localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, 1395–1403. New York, NY, USA: Association for Computing Machinery. ISBN 9781450379885.

Zhu, S.; Shah, M.; and Chen, C. 2022. TransGeo: Transformer Is All You Need for Cross-View Image Geo-Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1162–1171.