CVGL: Causal Learning and Geometric Topology

Songsong Ouyang **Yingying Zhu***
College of Computer Science and Software Engineering
Shenzhen University
2400101027@mails.szu.edu.cn, zhuyy@szu.edu.cn

Abstract

Cross-view geo-localization (CVGL) aims to estimate the geographic location of a street image by matching it with a corresponding aerial image. This is critical for autonomous navigation and mapping in complex real-world scenarios. However, the task remains challenging due to significant viewpoint differences and the influence of confounding factors. To tackle these issues, we propose the Causal Learning and Geometric Topology (CLGT) framework, which integrates two key components: a Causal Feature Extractor (CFE) that mitigates the influence of confounding factors by leveraging causal intervention to encourage the model to focus on stable, task-relevant semantics; and a Geometric Topology Fusion (GT Fusion) module that injects Bird's Eye View (BEV) road topology into street features to alleviate cross-view inconsistencies caused by extreme perspective changes. Additionally, we introduce a Data-Adaptive Pooling (DA Pooling) module to enhance the representation of semantically rich regions. Extensive experiments on CVUSA, CVACT, and their robustness-enhanced variants (CVUSA-C-ALL and CVACT-C-ALL) demonstrate that CLGT achieves state-of-the-art performance, particularly under challenging real-world corruptions. Our codes are available at CLGT.

1 Introduction

Cross-view geo-localization (CVGL) aims to estimate the geographic location of a street image by matching it to a corresponding aerial image. This task plays a crucial role in applications such as autonomous driving, robotic navigation, and urban mapping [24; 2; 27]. However, it remains highly challenging due to the extreme differences in perspective, scale, appearance, confounders and occlusion between street and aerial views. Previous studies have mainly explored three directions to improve cross-view matching: viewpoint modeling [28], spatial alignment [21], and hard negative mining [1]. Despite these efforts, cross-view geo-localization remains challenging due to weather changes, misalignment, and occlusions, all of which demand stronger generalization and discriminative feature learning. To address these limitations, Mi et al.[13] introduced feature consistency constraints to enhance robustness to orientation and field-of-view variations. To better reflect real-world conditions, Zhang et al.[32] proposed corruption-rich benchmarks for robust evaluation, while Ye et al.[28] leveraged a Bird's Eye View (BEV) representation as an intermediate domain to bridge the large cross-view gap.

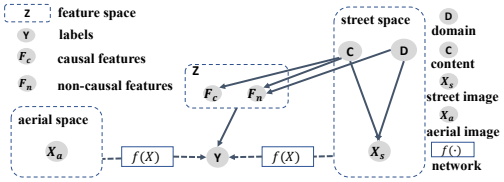

Figure 1: Structural Causal Model (SCM) for cross-view geo-localization. Nodes represent variables and arrows denote dependencies.

*Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

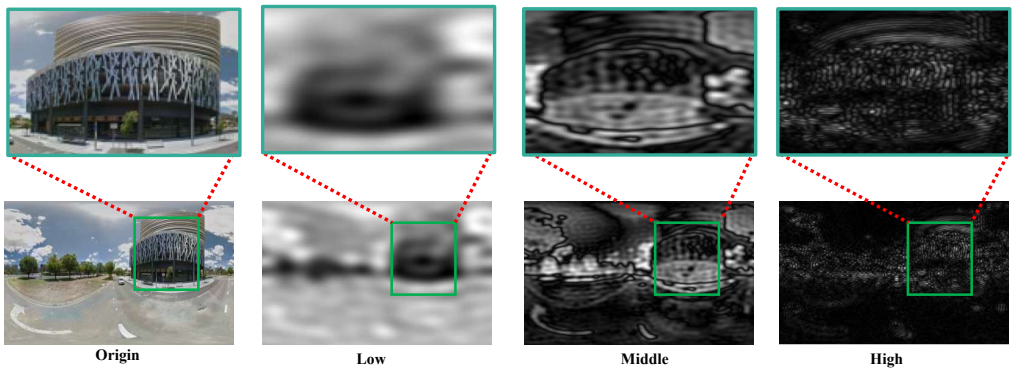

Figure 2: Visualization of low, mid, and high frequency components of a street image. Low and high frequencies emphasize domain-specific information such as style, while the mid-frequency band retains domain-invariant cues such as structure and shape.

Building upon the aforementioned challenges and recent advances, we propose a framework to address cross-view geo-localization from both causal perspectives and geometric. Inspired by instances where classification models mistakenly associate sheep with grass, we argue that the CVGL task should not rely on confounding factors such as background or lighting. To mitigate the interference of confounding (non-causal) factors and spurious correlations, while enhancing model generalization, we introduce causal learning concepts into the CVGL task. Drawing on previous causal modeling work and considering the characteristics of street images, we establish the first Structural Causal Model (SCM) for CVGL, and perform causal intervention guided by the SCM. Since domain-specific (non-causal) signals often reside in extreme low and high frequency bands, while mid-frequency components typically preserve structure-relevant, discriminative information [7] (Figure 2), we design our Causal Feature Extractor as a do-operation in the frequency domain. This allows us to implement a back-door adjustment -akin to causal interventions -that increases the model's attention to causal factors while reducing interference from non-causal factors. To further enhance the model's geometric awareness and mitigate the impact of large viewpoint gaps, we propose the Geometric Topology Fusion (GT Fusion) module, which robustly integrates BEV road topology into street features, leveraging clearer and more localized road topology compared to complex street images.

At the feature level, conventional pooling layers often fail to capture rich semantic cues. To address this, we propose a DA Pooling module that dynamically refines feature representations, enabling the model to capture more context-aware information across diverse scenes and viewpoints.

In summary, our main contributions are:

- We are the first to introduce causal learning concepts into CVGL tasks by applying causal interventions to latent confounding factors, thereby reducing their influence on feature learning. This mechanism enables the model to focus on causally relevant information, such as building structures and road layouts, leading to improved robustness and generalization in complex environments.

- We propose a GT Fusion module that enhances the model's ability to perceive geometric information, mitigating the issue of large viewpoint discrepancies in CVGL tasks and providing more robust localization performance for CVGL.

- We design a DA Pooling to extract rich semantic information and enhance semantic representations across different environments.

Our work highlights the importance of structural reasoning and causal robustness in bridging the cross-view domain gap, setting a new direction for future research in geo-localization.

2 Related Work

2.1 Cross-view Geo-localization

Contrastive Learning-based methods. Contrastive learning has been widely applied in cross-view geolocation tasks [1; 25; 26; 21; 28; 31]. It helps mitigate feature distribution discrepancies between different viewpoints. For example, ConGeo [13] leveraged both single-view and cross-view contrastive losses while incorporating view-specific augmentation strategies. This effectively extracts robust feature representations, enabling the model to maintain high matching accuracy despite viewpoint limitations and orientation deviations. Moreover, Sample4Geo [1] proposed a simplified yet effective contrastive learning framework with a symmetric InfoNCE [19] loss, which fully utilizes all negative samples to accelerate model convergence.

Incorporation of Geometric Information. To address the challenges posed by drastic viewpoint variations, some approaches focused on extracting geometric layout information or leveraging BEV images to enforce geometric consistency constraints. For instance, GeoDTR [33] employed a geometric layout extractor to learn spatial correlations between aerial and street features, preventing overfitting to low-level details. Similarly, EP-BEV [28] and HC-Net [21] integrated BEV representations into cross-view geolocation to bridge the substantial differences between views. EP-BEV utilized a dual-branch structure to impose geometric consistency constraints, while HC-Net [21] directly reformulated cross-view geolocation as an image alignment problem.

2.2 Causality in Computer Vision

This limitation underscores the motivation for causal inference in visual learning: relying solely on statistical correlations in data is insufficient for reliably predicting counterfactual outcomes and may amplify spurious associations. Causal inference, by modeling the underlying data-generating mechanisms, aims to isolate invariant causal factors and thereby enhance generalization to unseen domains and conditions. To mitigate the interference of non-causal features and extract invariant causal representations, causal mechanisms have been widely adopted in computer vision [18; 3; 17; 30; 6].

In cross-view geo-localization tasks, it is essential to first establish causal relationships and then apply causal inference—including interventional estimation and counterfactual analysis—to eliminate confounding contextual factors and domain shifts. This approach enhances model robustness against domain variations and weather conditions, which pose significant challenges in this task. Drawing on previous causal modeling work [12; 10] and considering the characteristics of street images, we formally define the causal relationships cross-view geo-localization as shown in Figure1, and employ interventional estimation to block the direct influence of confounders, significantly improving generalization. There are two common methods for causal intervention: front-door adjustment and back-door adjustment. Front-door adjustment is used when non-causal factors (confounders) are unobserved, requiring the introduction of an intermediate variable M to reduce the influence of confounding factors. When confounders are observable, back-door adjustment is used, where confounding factors are directly intervened to reduce their impact.

3 Method: CLGT

This paper proposes a novel framework for cross-view geo-localization. A multi-head attention-based fusion module, which robustly integrates BEV features into street features via cross-attention and dual dynamic fusion, enforces geometric consistency constraints. To enhance causal features in street representations while mitigating the interference of non-causal features, we employ causal inference-based estimation and intervention. Furthermore, we introduce a DA pooling module to refine the fused features with rich semantic information. The overall model architecture is shown in Figure 3. The following sections provide a detailed introduction to our proposed method.

3.1 Preliminary

BEV Generation. Various methods exist for generating BEV images, including geometry-based transformations [21; 28], Transformer-based [35], and diffusion-based methods [29]. To balance

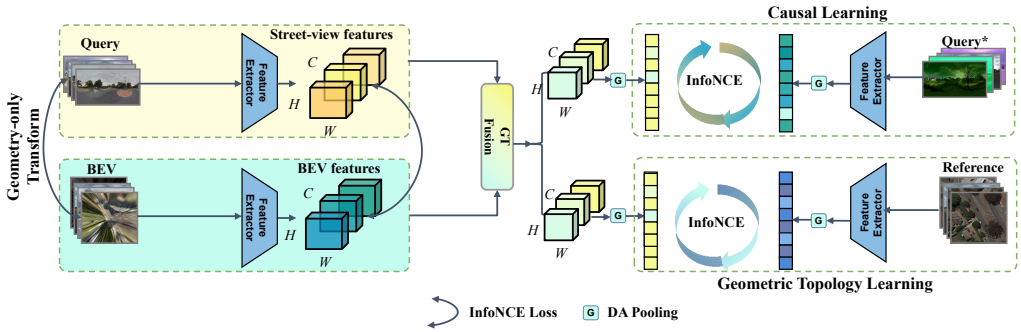

Figure 3: Overview of the proposed Causal Learning and Geometric Topology (CLGT) framework. The road topology information from BEV is fused via the GT Fusion module to obtain the fused features, which are then used for location matching with aerial image features. The causally enhanced street features from the CFE module provide causal supervision, and the DA Pooling module performs final feature extraction.

efficiency and memory use, we adopt the geometric transformation from [21], which directly computes BEV point positions via geometric back-projection from panoramic images. This explicit mapping projects street-view images into BEV space without relying on depth estimation or camera parameters, enabling a simple and efficient street-to-BEV transformation.

Structural Causal Model. As illustrated in Figure 1, our SCM is grounded on the following assumptions:

- The street image X_s is mainly generated from two sources: semantic content C and a domain confounder D (e.g., background, lighting), denoted as $C \rightarrow X_s \leftarrow D$.
- The content C contains both discriminative and non-discriminative parts. Together with D, the non-discriminative part contributes to the generation of non-causal features F_n via $D \rightarrow F_n \leftarrow C$. The CFE module perturbs a portion of these non-causal components.
- The causal features F_c are derived from the discriminative part of C via $C \rightarrow F_c$.
- The full feature representation $Z = \{F_c, F_n\}$ influences the final prediction Y via $Z \rightarrow Y$, where Y is the matching label.

The overall cross-view matching process can be expressed as $X_a \rightarrow f(X) \rightarrow Y \leftarrow f(X) \leftarrow X_s$, where X_a denotes the aerial image. In this context, the SCM formulation $Z \rightarrow Y$ is a causal abstraction of the model's computational path $X_s \rightarrow f(X) \rightarrow Y$.

3.2 Causal Learning

The complete process of Causal Learning is illustrated in the top-right corner of Figure 3, where $Query^*$ is obtained through the Causal Features Extractor.

Causal Features Extractor. As shown in Figure 2, roads and buildings in street-view images tend to occupy the mid-frequency spectrum, while style variations are concentrated in the high and low ends, respectively. This aligns with the nature of CVGL, where structural elements are crucial for localization, and view-specific cues often act as noise. To isolate task-relevant features, we leverage the Discrete Cosine Transform (DCT) in our Causal Feature Extractor. Then our Content-aware Mask (CaM) constructs three concentric circular masks with initial radii of r_1, r_2, and r_3, dividing the frequency spectrum into four regions. Unlike prior work [23] that used fixed spectral thresholds, these radii are linearly increased based on image gradient magnitude (via Sobel operator), so that images with stronger gradients preserve more mid-frequency components, enabling better retention of causal information. Larger radii correspond to stronger Gaussian perturbations in outer frequency bands. This allows the model to adaptively preserve mid-frequency, causal components while suppressing non-causal signals. The masked frequencies are then transformed back via inverse DCT. The entire

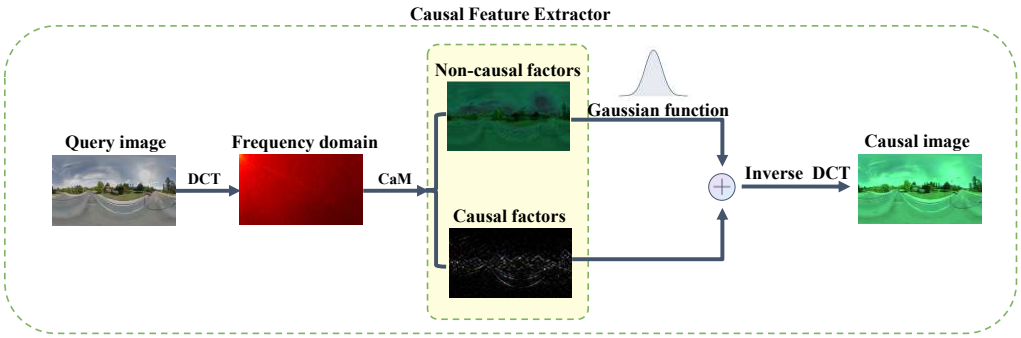

Figure 4: Illustration of the Causal Feature Extractor (CFE). The input image is transformed into the frequency domain via Discrete Cosine Transform (DCT). A Content-aware Mask (CaM) strategy dynamically separates mid-frequency causal components from low- and high-frequency non-causal components. A Gaussian function is applied to only to the non-causal components (e.g., lighting and brightness) to reduce their influence. Both causal and non-causal parts are then reconstructed via inverse (IDCT) to obtain the causally enhanced image.

process of CFE is shown in Figure 4. The Causal Features Extractor is defined as:

$$CFE(x) = \mathcal{F}' \left(\underbrace{(1 - M(r)) \cdot \mathcal{F}(x)}_{\text{Causal}} + \underbrace{\mathrm{G}(M(r) \cdot \mathcal{F}(x))}_{\text{Non-Causal Randomized}} \right) \tag{1}$$

where $\mathcal{F}$ denotes the Discrete Cosine Transform and $\mathcal{F}^{-1}$ is its inverse. $M(r)$ is a content-aware circular band-pass mask with radius r. $G(\cdot)$ denotes a randomization function, defined as $G(X) = X \cdot (1 + \mathcal{N}(0, 1))$.

After obtaining $Query^*$ through the CFE module (do$(X_s := X_s^*)$), we impose a supervision loss between the causally enhanced features derived from $Query^*$ and the fused features to weaken the path $C \rightarrow X_s \rightarrow f(X) \rightarrow Y$ (where C denotes confounding variables that influence the generation of X_s), achieving the effect similar to back-door adjustment in causal interventions.

3.3 Geometric Topology Learning

To guide the fusion of street and BEV features, we propose the GT Fusion module. This module effectively leverages the BEV road topology information while enriching the street features output by the backbone, dynamically injecting road topology information without compromising the street details.

GT Fusion. As shown in Figure 5 (left), our fusion module first applies a 3×3 depthwise convolution to backbone outputs $X_b, X_s \in \mathbb{R}^{C \times H \times W}$ to extract local features, maintaining the feature shape. To capture global context, instead of the common Spatial Reduction Attention (SRA) [11], which disrupts boundary spatial structure via non-overlapping token reduction, we adopt Overlapping Spatial Reduction (OSR) to preserve spatial coherence. Finally, X_s acts as the **query** and X_b as **key** and **value** in a cross-attention module, effectively fusing street-view and BEV features.

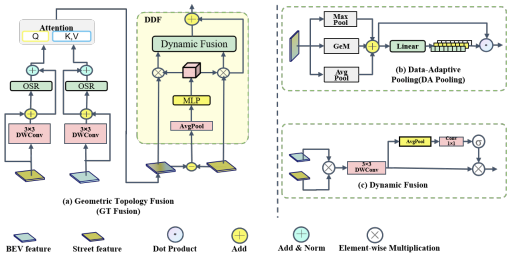

Figure 5: Overview of the GT Fusion and DA Pooling modules. GT Fusion uses cross-attention to exchange semantic information between street and BEV features, then uses Dual Dynamic Fusion (DDF) to enhance fusion robustness. DA Pooling employs a gating mechanism to adaptively weight features, highlighting the most informative ones.

Inspired by [4], we further introduce a Dual Dynamic Fusion (DDF) strategy to robustly integrate the original street features (denoted as $\mathbf{F}^{\mathrm{s}}$) and

their geometry-enhanced counterparts(denoted as $\mathbf{F}^{\mathrm{g}}$). DDF is defined as:

$$\begin{aligned} \mathbf{w} &= \sigma\left(\gamma\left(\text{AvgPool}(\mathbf{F}^{\mathrm{s}} + \mathbf{F}^{\mathrm{g}})\right)\right) \\ \text{Adaptive}(\mathbf{F}) &= \sigma(\mathbf{W} \cdot \text{AvgPool}(\mathbf{F})) \cdot \mathbf{F} \\ \mathbf{F}_{\text{fused}} &= \text{Adaptive}\left(\text{Conv}([\mathbf{w} \cdot \mathbf{F}^{\mathrm{s}},\ (1 - \mathbf{w}) \cdot \mathbf{F}^{\mathrm{g}}])\right) \end{aligned} \tag{2}$$

where $\mathbf{F}^{\mathrm{s}}$ denotes the original street-view feature. $\mathbf{W}$ is a 1×1 convolution, while Conv refers to a 3×3 convolution. $[\cdot, \cdot]$ denotes the concatenation operation along the channel dimension. γ is a linear transformation, and σ denotes the sigmoid activation function. The GT Fusion module can be formulated as follows:

$$X_s' = \text{proj}(X_s) + X_s, X_b' = \text{proj}(X_b) + X_b \tag{3}$$

$$Q = \text{LM}(\text{OSR}(X_s') + X_s'); K, V = \text{LM}(\text{OSR}(X_b') + X_b') \tag{4}$$

$$Z = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}} + B\right)V, \mathbf{F}_{\text{fused}} = \text{DDF}(X_s, Z) + X_s \tag{5}$$

where $proj$ refers to the 3×3 depthwise convolution, with X_s and X_b representing the street and BEV features output by the backbone, respectively. LM denotes a Layer Normalization. OSR stands for Overlapping Spatial Reduction, which is used for global information extraction. B is a relative position bias matrix that encodes the spatial relationships within the attention maps, d represents the number of channels in each attention head, and DDF denotes a dual dynamic fusion module.

Data-Adaptive Pooling. As shown on the right side of Figure 5, our Data-Adaptive Pooling module improves upon conventional pooling methods such as global max pooling, global average pooling, and adaptive pooling. These traditional pooling techniques, when used as the final feature aggregation step, fail to effectively capture the rich semantic information of the features. To enhance the representation capability, we combine global max pooling, global average pooling, and geometric mean (Gem) pooling into a single Gate Pooling module. This allows the model to autonomously learn the pooling output, and improve the quality of the final token, thus enhancing both the model's representational power and robustness, which can also be formulated as:

$$F_{max} = MaxPool(Z), F_{avg} = AvgPool(Z), F_{gem} = Gem(Z) \tag{6}$$

$$F_{output} = Gate(Linear(F_{max} + F_{avg} + F_{gem})) \tag{7}$$

where $MaxPool$ is global max pooling, $AvgPool$ is global average pooling, Gem denotes a geometric mean pooling, Linear is a Linear Functiion, and $Gate$ denotes a gating mechanism.

3.4 Loss Function

We apply InfoNCE loss between the fused features and the aerial image features, which serves as the primary supervision signal to optimize our model. The InfoNCE loss is defined as:

$$\mathcal{L}(f, S)_{\text{InfoNCE}} = -\log \frac{\exp(f \cdot r_+/\tau))}{\sum_{i=0}^{R} \exp(f \cdot r_i/\tau))} \tag{8}$$

where f denotes the fused feature guided by the query street image, and S is the set of encoded aerial images with one positive r_+ matching f. The InfoNCE loss computes the dot-product similarity between f and each r_i, maximizing similarity with r_+ and minimizing it with negatives. The temperature τ controls distribution sharpness and can be fixed or learned.

As stated in Section 3.2, we apply InfoNCE loss between the causally enhanced features and the fused features to achieve a similar effect to back-door adjustment, encouraging the fused features to focus on causal components. Prior to fusion, to encourage the BEV and street features to lie in a geometrically consistent space, we also apply InfoNCE loss between the two. To preserve their complementarity, we apply a scaling factor to control their learning balance. Thus, we obtain the overall loss of CLGT by computing the weighted sum of them as follows:

$$\mathcal{L}_{CLGT} = \mathcal{L}(f, S)_{infoNCE} + \gamma\mathcal{L}(f, s^*)_{infoNCE} + \alpha\mathcal{L}(s, b)_{infoNCE} \tag{9}$$

where α and γ are scaling coefficients, s^* denotes the causally enhanced street features, s represents the original street features, and b is the BEV feature.

4 Experiment

In our evaluation we conduct experiments on four standard benchmarks, namely CVUSA [22], CVACT [8], VIGOR [36] and CVACT_val-C-ALL, CVACT_test-C-ALL, CVUSA-C-ALL [32]. In the subsequent tables we compare our approach with previous work.

4.1 Dataset and Evaluation Protocol

Dataset. We evaluate our model on three widely-used cross-view geo-localization benchmarks—CVUSA, CVACT, and VIGOR—as well as their robust variants: CVACT_val-C-ALL, CVACT_test-C-ALL, and CVUSA-C-ALL, which introduce various real-world corruptions to test model robustness under challenging conditions. CVUSA and CVACT each provide 35,532 training and 8,884 testing image pairs with a strict 1-to-1 ground-to-aerial correspondence. In addition, CVACT offers an extra 92,802 GPS-tagged query images for large-scale retrieval evaluation, making it suitable for both standard and large-scale testing scenarios. VIGOR is a more challenging benchmark that spans four metropolitan areas—New York, Seattle, San Francisco, and Chicago—and includes 105,214 query and 90,618 reference images. Unlike CVUSA and CVACT, VIGOR introduces a harder retrieval setup by assigning each query one true positive and three semi-positive samples, thus increasing the difficulty of discriminative matching. It also supports both same-city and cross-city evaluation settings to assess generalization. To further assess model robustness in realistic conditions, we employ corruption-augmented datasets: CVACT_val-C-ALL, CVACT_test-C-ALL, and CVUSA-C-ALL. These variants simulate 16 types of visual degradations, generating approximately 1.5 million corrupted images in total. They provide a rigorous benchmark for evaluating the model's ability to maintain performance under various environmental and sensor-induced perturbations.

Evaluation Protocol. We adopt Recall@K as the primary evaluation metric, where $K \in \{1, 5, 10\}$, as well as Recall@1%. A query is considered correctly localized if its corresponding aerial image appears among the top-K retrieved candidates for a given street panorama.

4.2 Implementation Details

During the retrieval stage, we adopt ConvNeXt-B as the backbone encoder for both street images and aerial images. Our baseline is EP-BEV. To reduce computational cost and memory usage, we set the image resolution to 384×384, consistent with EP-BEV. The model is optimized using AdamW with an initial learning rate of 0.5×10^{-3}. We train the network for 40 epochs with a batch size of 128. The training is conducted on eight 32GB NVIDIA V100 GPUs. For both α and γ in Equation 9, we set their values to 0.1 to provide auxiliary supervision without overwhelming the main optimization objective. When we increase the value of γ, the model performance improves across various datasets. However, to prevent the value from becoming too large and causing model collapse, which would negatively affect the matching between street and aerial images, we set a default value of 0.1, although this collapse was not observed during training. The optimal value is 0.5, and we will also provide hyperparameter experiments and model performance with $\gamma = 0.5$ in the supplementary materials. We set the initial three radii for the content-aware mask to 0.1, 0.3, and 0.6, respectively. We also observe that performance is stable under small variations in the initial radius. Other training settings follow those used in Sample4Geo.

Table 1: Ablation study on causal learning: comparisons of performance on the CVACT_val-C-ALL and CVACT_test-C-ALL datasets.

Model	CVACT_val-C-ALL			CVACT_test-C-ALL		
	R@1	R@5	R@10	R@1	R@5	R@10
Ours	**88.68**	**95.58**	**96.66**	**69.06**	**90.12**	**92.70**
Only CL	88.11	94.91	96.04	67.88	89.31	91.85
Baseline	85.94	94.52	95.93	64.62	87.75	90.78

4.3 Comparing with State-of-the-art Models

Cross-view Image retrieval. As shown in Table 2, our method achieves the best overall performance across CVUSA, CVACT_val, and CVACT_test datasets. Compared with the strong baseline EP-BEV, our model improves Recall@1 from 97.41% to 98.85% on CVUSA, and from 90.61% to 91.97% on CVACT_val. On the more realistic CVACT_test set, we achieve 73.22% Recall@1, surpassing EP-BEV by 1.81% points.These consistent gains demonstrate the effectiveness of our design. The

integration of BEV-based geometric topology helps capture structured layout cues, while the causal learning strategy improves robustness by suppressing spurious visual signals. Together, they enable more discriminative and generalizable representations for cross-view geo-localization. Results on the VIGOR dataset are provided in the supplementary materials.

Table 2: Comparisons with state-of-the-art models on the CVUSA, CVACT_val and CVACT_test datasets. (†methods that use polar transformation.)

Model	CVUSA				CVACT_val				CVACT_test			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
SAFA† [14]	89.84	96.93	98.14	99.64	81.03	92.80	94.84	-	-	-	-	-
LPN [20]	85.79	95.38	96.98	99.41	79.99	90.63	92.56	-	-	-	-	-
LPN† [20]	92.83	98.00	98.85	99.78	83.66	94.14	95.92	98.41	-	-	-	
DSM [16]	91.96	97.50	98.54	99.67	82.49	92.44	93.99	97.32	-	-	-	
TransGeo [37]	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	
GeoDTR [33]	93.76	98.47	99.22	99.85	85.43	94.81	96.11	98.26	62.96	87.35	90.70	98.61
GeoDTR+ [34]	95.05	98.42	98.92	99.77	87.76	95.50	96.50	98.32	67.75	90.15	92.73	98.53
GeoDTR† [33]	95.43	98.86	99.34	99.86	86.21	95.44	96.72	98.77	64.52	88.59	91.96	98.74
Sample4G [1]	98.68	99.68	99.78	99.87	90.81	96.74	97.48	98.77	71.51	92.42	94.45	98.70
ConGeo [13]	98.30	-	-	**99.90**	90.10	-	-	98.20	71.70	98.30	-	-
EP-BEV [28]	97.41	99.40	99.60	99.76	90.61	96.57	97.32	98.71	71.41	92.38	94.37	98.77
Ours	98.73	**99.71**	99.80	99.84	91.61	96.93	**97.72**	**98.77**	73.03	93.03	94.81	98.63
Ours ($\gamma = 0.5$)	**98.85**	**99.71**	**99.81**	99.86	**91.97**	**96.95**	**97.72**	**98.77**	**73.22**	**93.50**	**95.23**	**98.79**

Robustness Evaluation. As shown in Table 3, our method consistently outperforms baselines across robust datasets, with an average improvement of 5.00%. Notably, it achieves a 6.62% gain on CVUSA-C-ALL, highlighting the model's ability to extract localization-relevant cues such as edge textures of buildings and road structures, while suppressing non-causal noise like lighting and weather conditions. On challenging splits such as CVACT_val-C-ALL, CVACT_test-C-ALL, and CVUSA-C-ALL, our approach demonstrates strong robustness by mitigating the impact of 16 common perturbations and improving retrieval accuracy. Furthermore, in cross-dataset evaluation (trained on CVUSA, tested on CVACT), our method improves performance by 5.50% and 2.84% (Table 6), further validating its generalization and robustness under distribution shifts.

Table 3: Comparisons with state-of-the-art models on the CVUSA-C-ALL, CVACT_val-C-ALL and CVACT_test-C-ALL datasets.

Model	CVUSA-C-ALL				CVACT_val-C-ALL				CVACT_test-C-ALL			
	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%	R@1	R@5	R@10	R@1%
CVM-Net [5]	6.09	16.05	23.14	52.51	-	-	-	-	-	-	-	-
OriCNN [9]	9.38	22.26	30.04	58.99	15.31	28.31	35.21	58.39	3.69	8.33	11.04	43.93
SAFA [14]	63.68	78.08	82.82	93.91	56.72	73.60	78.59	91.32	31.18	52.06	58.60	90.41
CVFT [15]	41.05	64.01	72.64	91.37	45.69	66.45	72.97	88.38	22.82	43.48	51.07	88.99
DSM [16]	75.27	86.26	89.42	95.07	70.04	82.81	85.86	93.51	47.13	68.41	73.52	93.18
L2LTR [25]	87.93	95.45	97.01	99.01	82.13	93.34	94.93	98.10	57.20	82.59	87.23	98.09
TransGeo [37]	82.72	91.95	94.03	97.92	74.04	86.19	89.10	94.98	52.18	74.35	78.99	95.03
GeoDTR [33]	84.64	93.29	95.01	98.24	77.40	88.95	91.28	95.91	52.87	78.84	83.17	95.84
EP-BEV [28]	86.22	94.86	96.58	99.00	85.94	94.52	95.93	98.21	64.62	87.75	90.78	98.43
Ours	92.64	97.21	98.21	**99.35**	88.68	95.58	96.66	**98.49**	69.06	90.12	92.70	98.41
Ours ($\gamma = 0.5$)	**92.84**	**97.61**	**98.41**	**99.35**	**89.49**	**95.84**	**96.92**	**98.49**	**69.71**	**91.05**	**93.30**	**98.85**

Table 4: Ablation study on DA Pooling: comparison with other pooling methods on CVACT.

Method	CVACT_val			CVACT_test		
	R@1	R@5	R@10	R@1	R@5	R@10
DA pooling	**91.61**	**96.93**	**97.72**	**73.03**	**93.03**	**94.81**
Gem	89.24	95.64	96.65	69.78	92.18	93.78
Avg	91.19	96.52	97.33	71.58	92.52	94.51
Max	90.55	96.51	97.25	70.78	92.20	94.23

Table 5: Ablation study of the CLGT on CVACT_val.

Method	CVACT_val			
	R@1	R@5	R@10	R@1%
CLGT	**91.61**	**96.93**	**97.72**	**98.77**
w/o GT Fusion	91.24	96.77	97.41	98.65
w/o CFE	91.01	96.66	97.51	98.64
Baseline	90.61	96.57	97.32	98.71

Ablation Study. We conduct comprehensive ablation studies on CVACT_val, CVACT_val-C-ALL, and CVACT_test-C-ALL to evaluate the individual contributions of each proposed component. As reported in Table 5, employing only the Geometric Topology Fusion (GT Fusion) module

yields a Recall@1 of 91.01% on CVACT_val. This result highlights the importance of geometric consistency learning, which enables the model to better capture road layouts and spatial structures, thereby improving the alignment between ground-level and aerial images.When using only the CFE module, the model achieves a Recall@1 of 91.24%, illustrating its capability to suppress spurious correlations and guide the model toward learning causally relevant and semantically stable features. This enhancement plays a crucial role in resisting interference from latent confounders. As shown in Table 4, our DA Pooling consistently outperforms traditional pooling schemes across all evaluation metrics. Specifically, compared to GeM, DA Pooling improves Recall@1 on CVACT by +2.37%, showing its advantage in dynamically emphasizing informative spatial regions. This confirms the importance of adaptivity in multi-view feature aggregation.

Table 6: Results on different cross-view transfer tasks. Each task is evaluated with representative methods. (†Methods using polar transformation.)

Task	Method	R@1	R@5	R@10	R@1%
CVUSA → CVACT_val	L2LTR [25]	47.55	70.58	-	91.39
	L2LTR† [25]	52.58	75.81	-	93.51
	GeoDTR [33]	47.79	70.52	-	92.20
	GeoDTR† [33]	53.16	75.62	-	93.80
	Samp4G [1]	56.62	77.79	87.02	94.69
	EP-BEV [28]	59.32	80.79	86.02	94.69
	Ours	60.70	81.40	86.10	95.16
	Ours ($\gamma = 0.5$)	**64.82**	**84.38**	**88.77**	**96.16**
CVUSA → CVACT_test	L2LTR [25]	-	-	-	-
	L2LTR† [25]	-	-	-	-
	GeoDTR [33]	11.24	18.69	23.67	72.09
	GeoDTR† [33]	22.09	32.22	39.59	85.53
	Samp4G [1]	27.78	52.08	60.33	94.88
	EP-BEV [28]	32.68	58.62	65.34	95.21
	Ours	33.23	59.59	67.53	95.31
	Ours ($\gamma = 0.5$)	**35.52**	**63.37**	**71.40**	**96.35**

Furthermore, as shown in Table 1, causal learning alone brings significant improvements on the more challenging and corrupted test sets: Recall@1 increases by 2.23% on CVACT_val-C-ALL and by 3.20% on CVACT_test-C-ALL. In addition to Recall@1, other evaluation metrics such as Recall@5 and Recall@10 also show consistent improvements, closely approaching the performance of the full CLGT model. These results confirm the effectiveness of our causal learning strategy in improving robustness under real-world corruptions and diverse input conditions. Overall, the ablation results validate that both modules—GT Fusion and CFE—contribute meaningfully and complement each other in addressing the challenges of cross-view geo-localization.

Visualization Analysis. To qualitatively assess the effectiveness of CLGT, we visualize attention heatmaps generated by the baseline and CLGT models on test images from the CVUSA dataset. As shown in Figure 6, we first visualize the heatmaps on clean images. It can be seen that the baseline model's attention is more scattered, even focusing on the sky and other background noise, while CLGT consistently attends to task-relevant information, especially road structures. Under heavy snow conditions, compared to the baseline, the regions attended by our model remain almost unchanged, whereas the baseline's focus is completely misaligned. This shows that our CLGT consistently attends to semantically meaningful structures, such as road intersections and corner layouts, which are more stable across views. This demonstrates the effectiveness of our design in guiding the model to prioritize task-relevant features and suppress distractions, leading to improved cross-view discriminability. Visualizations for other corruption types can be found in the supplementary materials.

Complexity Analysis. As shown in Table 7, we report both GFLOPs and average inference time (in milliseconds per batch of 128 images) on the CVACT_val set. The results demonstrate that our method achieves the best R@1 accuracy with only marginal computational overhead compared to other methods.

Table 7: Comparison of GFLOPs and average inference time per batch (batch size = 128) on CVACT_val. Avg inference time (ms) represents the mean time to process one batch.

Method	GFLOPs	Avg Inference Time (ms)	CVACT_val R@1
Sample4Geo	**90.54**	**2367.55**	90.81
EP-BEV	**90.54**	2396.53	90.61
Ours	90.56	2374.76	**91.61**

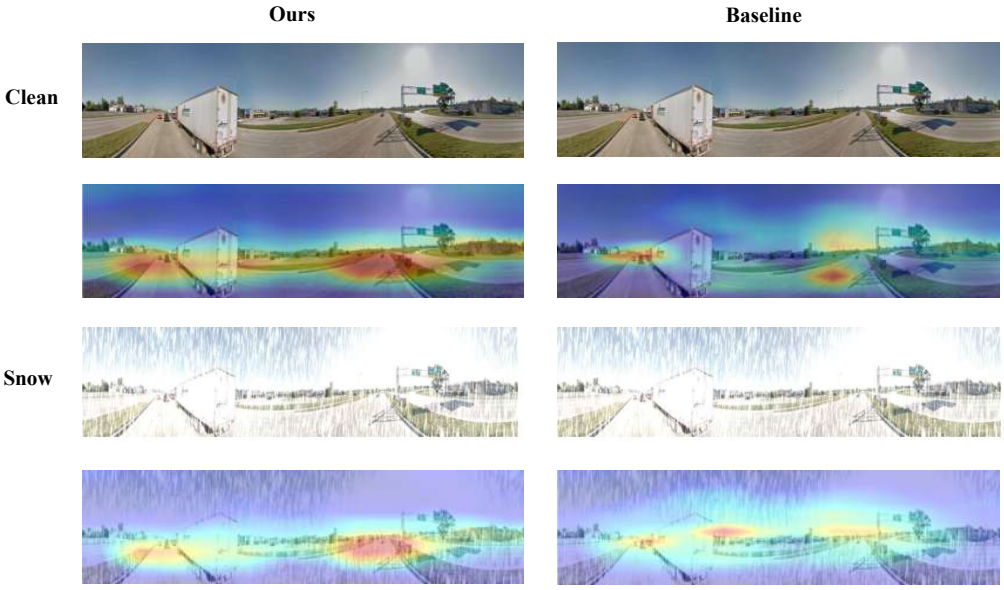

Figure 6: Heatmap visualizations on CVUSA under clean and snow settings. Compared to the baseline, our method focuses more on task-relevant structural information. Under heavy snow conditions, our method remains highly robust, with attention regions largely unchanged, whereas the baseline's attention is completely misaligned.

5 Conclusion and Future Work

In this work, we present a novel cross-view geo-localization framework that integrates BEV-street view fusion with causal learning mechanism. Unlike previous methods that utilize BEV merely as an auxiliary representation, our approach enables feature-level interaction that effectively and robustly incorporate road topology. To further improve generalization, we introduce a causal intervention module, thereby enhancing filters out non-causal information and enhances model robustness under various conditions. Experimental results on both standard and challenging datasets demonstrate consistent performance gains. Nonetheless, the BEV representations derived from geometric transformations contains considerable noise, which limits further the performance improvements. Future work will explore more advanced causal inference strategies tailored to the dynamics of complex cross-view localization tasks.

6 Acknowledgments

This work was supported in part by Shenzhen Science and Technology Program under Grant JCYJ20240813142510014 and Grant 20220810142553001, in part by the Key Project of Department of Education of Guangdong Province under Grant 2023ZDZX1016, and in part by the National Natural Science Foundation of China under Grant 62072318 and Grant U22A2097.

References

[1] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation, 2023. URL https://arxiv.org/abs/2303.11851.

[2] Jiqi Fan, Enhui Zheng, Yufei He, and Jianxing Yang. A cross-view geo-localization algorithm using uav image and satellite image. *Sensors*, 24(12), 2024. ISSN 1424-8220. doi: 10.3390/s24123719. URL https://www.mdpi.com/1424-8220/24/12/3719.

[3] Yash Goyal, Amir Feder, Uri Shalit, and Been Kim. Explaining classifiers with causal concept effect (cace), 2020. URL https://arxiv.org/abs/1907.07165.

[4] Xiaoshuai Hao, Yunfeng Diao, Mengchuan Wei, Yifan Yang, Peng Hao, Rong Yin, Hui Zhang, Weiming Li, Shu Zhao, and Yu Liu. Mapfusion: A novel bev feature fusion network for multi-modal map construction, 2025. URL https://arxiv.org/abs/2502.04377.

[5] Sixing Hu, Mengdan Feng, Rang M. H. Nguyen, and Gim Hee Lee. Cvm-net: Cross-view matching network for image-based ground-to-aerial geo-localization. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7258–7267, 2018. doi: 10.1109/CVPR.2018.00758.

[6] Xinting Hu, Kaihua Tang, Chunyan Miao, Xian-Sheng Hua, and Hanwang Zhang. Distilling causal effect of data in class-incremental learning, 2021. URL https://arxiv.org/abs/2103.01737.

[7] Jiaxing Huang, Dayan Guan, Aoran Xiao, and Shijian Lu. Fsdr: Frequency space domain randomization for domain generalization. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6887–6898, 2021. doi: 10.1109/CVPR46437.2021.00682.

[8] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5617–5626, 2019. doi: 10.1109/CVPR.2019.00577.

[9] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization, 2019. URL https://arxiv.org/abs/1903.12351.

[10] Yajing Liu, Shijun Zhou, Xiyao Liu, Chunhui Hao, Baojie Fan, and Jiandong Tian. Unbiased faster r-cnn for single-source domain generalized object detection, 2024. URL https://arxiv.org/abs/2405.15225.

[11] Meng Lou, Shu Zhang, Hong-Yu Zhou, Sibei Yang, Chuan Wu, and Yizhou Yu. Transxnet: Learning both global and local dynamics with a dual dynamic token mixer for visual recognition. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–14, 2025. doi: 10.1109/TNNLS.2025.3550979.

[12] Fangrui Lv, Jian Liang, Shuang Li, Bin Zang, Chi Harold Liu, Ziteng Wang, and Di Liu. Causality inspired representation learning for domain generalization, 2022. URL https://arxiv.org/abs/2203.14237.

[13] Li Mi, Chang Xu, Javiera Castillo-Navarro, Syrielle Montariol, Wen Yang, Antoine Bosselut, and Devis Tuia. Congeo: Robust cross-view geo-localization across ground view variations, 2024. URL https://arxiv.org/abs/2403.13965.

[14] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/ba2f0015122a5955f8b3a50240fb91b2-Paper.pdf.

[15] Yujiao Shi, Xin Yu, Liu Liu, Tong Zhang, and Hongdong Li. Optimal feature transport for cross-view image geo-localization, 2019. URL https://arxiv.org/abs/1907.05021.

[16] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where Am I Looking At? Joint Location and Orientation Estimation by Cross-View Matching . In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4063–4071, Los Alamitos, CA, USA, June 2020. IEEE Computer Society. doi: 10.1109/CVPR42600.2020.00412. URL https://doi.ieeecomputersociety.org/10.1109/CVPR42600.2020.00412.

[17] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. Long-tailed classification by keeping the good and removing the bad momentum causal effect, 2021. URL https://arxiv.org/abs/2009.12991.

[18] Yingjie Tian, Kunlong Bai, Xiaotong Yu, and Siyu Zhu. Causal multi-label learning for image classification. *Neural Networks*, 167:626–637, 2023. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.08.052. URL https://www.sciencedirect.com/science/article/pii/S0893608023004732.

[19] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. URL http://arxiv.org/abs/1807.03748.

[20] Tingyu Wang, Zhedong Zheng, Chenggang Yan, Jiyong Zhang, Yaoqi Sun, Bolun Zheng, and Yi Yang. Each part matters: Local patterns facilitate cross-view geo-localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(2):867–879, February 2022. ISSN 1558-2205. doi: 10.1109/tcsvt.2021.3061265. URL http://dx.doi.org/10.1109/TCSVT.2021.3061265.

[21] Xiaolong Wang, Runsen Xu, Zuofan Cui, Zeyu Wan, and Yu Zhang. Fine-grained cross-view geo-localization using a correlation-aware homography estimator, 2023. URL https://arxiv.org/abs/2308.16906.

[22] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery, 2015. URL https://arxiv.org/abs/1510.03743.

[23] Mingjun Xu, Lingyun Qin, Weijie Chen, Shiliang Pu, and Lei Zhang. Multi-view adversarial discriminator: Mine the non-causal factors for object detection in unseen domains, 2023. URL https://arxiv.org/abs/2304.02950.

[24] Yuanze Xu, Ming Dai, Wenxiao Cai, and Wankou Yang. Precise gps-denied uav self-positioning via context-enhanced cross-view geo-localization, 2025. URL https://arxiv.org/abs/2502.11408.

[25] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29009–29020. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/f31b20466ae89669f9741e047487eb37-Paper.pdf.

[26] Zhanyuan Yang, Jinghua Wang, and Yingying Zhu. Few-shot classification with contrastive learning, 2022. URL https://arxiv.org/abs/2209.08224.

[27] Yuwen Yao, Cheng Sun, Tao Wang, Jianxing Yang, and Enhui Zheng. Uav geo-localization dataset and method based on cross-view matching. *Sensors*, 24(21), 2024. ISSN 1424-8220. doi: 10.3390/s24216905. URL https://www.mdpi.com/1424-8220/24/21/6905.

[28] Junyan Ye, Zhutao Lv, Weijia Li, Jinhua Yu, Haote Yang, Huaping Zhong, and Conghui He. Cross-view image geo-localization with panorama-bev co-retrieval network, 2024. URL https://arxiv.org/abs/2408.05475.

[29] Xin Ye, Burhaneddin Yaman, Sheng Cheng, Feng Tao, Abhirup Mallik, and Liu Ren. Bevdiffuser: Plug-and-play diffusion model for bev denoising with ground-truth guidance, 2025. URL https://arxiv.org/abs/2502.19694.

[30] Zhongqi Yue, Hanwang Zhang, Qianru Sun, and Xian-Sheng Hua. Interventional few-shot learning. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 2734–2746. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1cc8a8ea51cd0adddf5dab504a285915-Paper.pdf.

[31] Qingwang Zhang and Yingying Zhu. Aligning geometric spatial layout in cross-view geo-localization via feature recombination. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN 978-1-57735-887-9. doi: 10.1609/aaai.v38i7.28554. URL https://doi.org/10.1609/aaai.v38i7.28554.

[32] Qingwang Zhang, hongji yang, and Yingying Zhu. Benchmarking the robustness of cross-view geo-localization models, 2024. URL https://openreview.net/forum?id=x8mzNomCRe.

[33] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning disentangled geometric layout correspondence, 2023. URL https://arxiv.org/abs/2212.04074.

[34] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: Toward generic cross-view geolocalization via geometric disentanglement, 2024. URL https://arxiv.org/abs/2308.09624.

[35] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation, 2022. URL https://arxiv.org/abs/2205.02833.

[36] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5316–5325, 2021. doi: 10.1109/CVPR46437.2021.00364.

[37] Sijie Zhu, Mubarak Shah, and Chen Chen. Transgeo: Transformer is all you need for cross-view image geo-localization. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1152–1161, 2022. doi: 10.1109/CVPR52688.2022.00123.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims in the abstract and introduction align well with the methodologies, experiments, and findings presented in the paper.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We discussed the limitations of the paper in the conclusion section.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Each theoretical result is accompanied by clearly stated assumptions and complete, rigorous proofs.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Thi paper provides detailed descriptions of the experimental setup, model architecture, training procedures, and evaluation metrics necessary to reproduce the main results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will provide the code.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have specified all training and testing details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: We have not conducted experiments with error bars yet.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: We have provided detailed information on the computing resources used for each experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research we conducted fully conforms to the NeurIPS Code of Ethics in all respects.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses both the potential positive and negative societal impacts of the work performed, addressing its broader implications responsibly.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not involve any data or models that pose high risks for misuse.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have properly credited the creators and original owners of all assets used in the paper, and explicitly mentioned and respected their licenses and terms of use.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [NA]

Justification: The paper does not release any new assets.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.