Geo²: Geometry-Guided Cross-view Geo-Localization and Image Synthesis

Yancheng Zhang[1] Xiaohan Zhang[2] Guangyu Sun[1] Zonglin Lyu[1] Safwan Wshah[2] Chen Chen[1]
[1]Institute of Artificial Intelligence, University of Central Florida, [2]University of Vermont
{yczhang,guangyu.sun,zonglin.lyu,chen.chen}@ucf.edu, {xiaohan.zhang,safwan.wshah}@uvm.edu

Abstract

Cross-view geo-spatial learning consists of two important tasks: Cross-View Geo-Localization (CVGL) and Cross-View Image Synthesis (CVIS), both of which rely on establishing geometric correspondences between ground and aerial views. Recent Geometric Foundation Models (GFMs) have demonstrated strong capabilities in extracting generalizable 3D geometric features from images, but their potential in cross-view geo-spatial tasks remains underexplored. In this work, we present ***Geo²****, a unified framework that leverages* ***Geo****metric priors from GFMs (e.g., VGGT) to jointly perform* ***Geo****-spatial tasks, CVGL and bidirectional CVIS. Despite the 3D reconstruction ability of GFMs, directly applying them to CVGL and CVIS remains challenging due to the large viewpoint gap between ground and aerial imagery. We propose GeoMap, which embeds ground and aerial features into a shared 3D-aware latent space, effectively reducing cross-view discrepancies for localization. This shared latent space naturally bridges cross-view image synthesis in both directions. To exploit this, we propose GeoFlow, a flow-matching model conditioned on geometry-aware latent embeddings. We further introduce a consistency loss to enforce latent alignment between the two synthesis directions, ensuring bidirectional coherence. Extensive experiments on standard benchmarks, including CVUSA, CVACT, and VIGOR, demonstrate that Geo² achieves state-of-the-art performance in both localization and synthesis, highlighting the effectiveness of 3D geometric priors for cross-view geo-spatial learning. Our source code can be accessed through https://fobow.github.io/geo2.github.io/.*

1. Introduction

Cross-view geo-spatial learning primarily comprises Cross-View Geo-Localization (CVGL) and Cross-View Image Synthesis (CVIS). CVGL aims to determine the location of a query ground image by matching it against a database of geo-tagged images. CVIS, on the other hand, aims to synthesize a corresponding view from another view, such as generating a satellite view from a ground view (Ground-to-Satellite or G2S) or a ground view from a satellite view (Satellite-to-Ground or S2G). In summary, both tasks focus on learning consistent feature representations across different views, facilitating more accurate geo-localization and higher-quality image synthesis.

A key challenge in cross-view geo-spatial tasks is the significant appearance and geometric gap between images from different views, such as ground and satellite perspectives. Extracting and aligning geometric characteristics from these views is therefore crucial. For example, in CVGL, GeoDTR [50, 51] introduces a geometric layout extractor to capture structural information from both ground and satellite images, leading to notable performance improvements. Similarly, in CVIS, many methods incorporate explicit geometric cues, including height estimation [21], geometric projection [45], and volume density modeling [26]. These geometric features effectively enhance the quality of the synthesized image pairs.

Despite the fact that both CVGL and CVIS benefit from geometric guidance, most existing works treat these two tasks separately. Furthermore, the geometry information used in these methods often relies on customized modules [50, 51] or predefined geometric transformations [29]. While such components can capture geometric hints to some extent, they are typically tailored for individual tasks, which limits their generalizability. As a result, CVGL and bi-directional CVIS are rarely able to benefit from each other in a unified framework. For example, BEV estimation has been shown to improve ground-to-satellite synthesis [2], but it is difficult to extend this approach to the reverse direction. A more generalizable geometric prior is therefore crucial for allowing CVGL and bidirectional CVIS to mutually enhance each other.

Recently, Geometric Foundation Models (GFMs) such as DUSt3R [37], MASt3R [14], and VGGT [35] have demonstrated strong capabilities in 3D understanding and reconstruction. These models can extract generalizable geometric attributes from multi-view or even single-view images, making them highly adaptable to diverse visual settings. AerialMegaDepth [34] further fine-tunes MASt3R

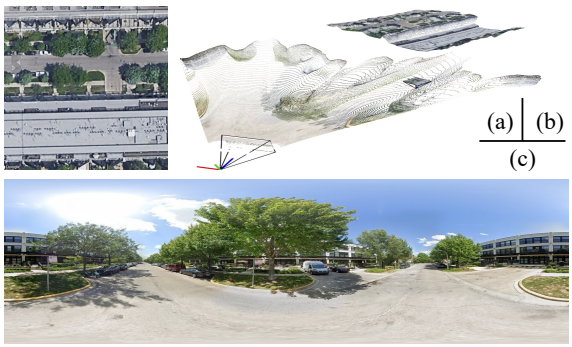

Figure 1. Illustration of directly using VGGT on satellite (a) and ground (c) images, leading to incorrect reconstructed shown in (b).

and DUSt3R on aerial and ground imagery, improving cross-view reconstruction in the challenging aerial-ground domain. This suggests that GFMs are not only effective for traditional 3D tasks, but also hold promise for cross-view applications. However, the potential of GFMs in cross-view geo-spatial learning remains largely underexplored. As shown in Figure 1, effectively leveraging GFMs for cross-view geo-spatial tasks is non-trivial. For example, naïvely applying VGGT to cross-view image pairs often results in inaccurate geometry, due to the substantial viewpoint differences between ground and satellite imagery.

To address the above-mentioned challenges, we propose Geo^2, a geometry-guided framework that integrates 3D priors into cross-view geo-spatial tasks. The key idea of Geo^2 is to embed ground and satellite images into a shared geometry-aware latent space, which incorporates geometry priors from VGGT and bridges CVGL and CVIS, enabling consistent cross-view understanding and generation. Geo^2 supports joint learning of CVGL and bidirectional CVIS, *i.e.*, ground-to-aerial and aerial-to-ground, **without re-training**. Specifically, for the CVGL task, we propose GeoMap, a dual-branch model that encodes ground and satellite images separately into geometry-aware features. Given the inherent similarity between CVGL and CVIS, these geometry-aware features naturally facilitate the geometric consistency for the CVIS task. Accordingly, we propose GeoFlow for bidirectional image synthesis. Our contributions can be summarized as follows:

- We propose Geo^2, a novel framework that incorporates 3D priors from Geometric Foundation Models to jointly perform Cross-View Geo-Localization (CVGL) and bidirectional Cross-View Image Synthesis (CVIS).
- We introduce GeoMap, a dual-branch model that aligns ground and satellite views in a shared geometry-aware latent space, improving cross-view embedding alignment for both localization and image synthesis.
- We propose GeoFlow, a flow-matching model that naturally supports bi-directional generation, conditioned on the geometry-aware features from GeoMap. We further introduce a consistency loss to improve geometric consistency between the two synthesis directions.
- We conduct extensive experiments on multiple cross-view benchmarks [18, 41, 54], demonstrating that Geo^2 achieves outstanding performance on both localization and synthesis tasks, validating the effectiveness of our geometry-guided design.

Table 1. Comparison of representative cross-view geo-spatial learning methods on supported tasks and geometric priors.

Method	CVGL	CVIS	Bi-dir	Joint	Geometric Prior
GeoDTR+ [51]	✓	✗	✗	✗	GLE [51]
Sample4Geo [8]	✓	✗	✗	✗	✗
RGCIS [44]	✗	✓	✓	✗	SAIG [56]
Sat2Density [26]	✗	✓	✗	✗	Volume Density
CDE [33]	✓	✓	✗	✓	SAFA [29]
Ours	✓	✓	✓	✓	VGGT [35]

2. Related Work

Cross-View Geo-Spatial Learning. The goal of Cross-View Learning is to model the correlation between the ground view (i.e., street-view images) and the overhead view (i.e., satellite imagery). There are mainly two tasks, Cross-View Image Synthesis (CVIS) and Cross-View Geo-Localization (CVGL) [40]. CVGL aims to localize a query ground image by matching against a geo-tagged satellite database. Prior CVGL studies [29, 30, 33, 46, 50] attempt to tackle the drastic view changes by adopting polar transformation, which assumes center-alignment between satellite and ground images. To break such a strong prior, recent methods [43, 51, 55] have investigated extracting cross-view geometric correspondence features by using an attention mechanism. Moreover, several works [3, 8, 54] also adopt hard sample mining mechanisms to further guide the model to differentiate similar visual contextual information in different locations.

CVIS, on the other hand, targets to synthesize one view from another view (i.e., Aerial-to-Ground or A2G), or vice versa (G2A). Earlier studies explored additional semantic priors to enhance the structure of ground view synthesis results [28, 32, 42, 47]. More recent works further leveraged the auxiliary information, such as height and depth estimation [21, 31], volume estimation [15, 26], BEV estimation [2, 45], and the help of CVGL models [33, 44], to tackle this challenging problem. However, most of the prior studies are not inherently invertible, preventing these methods to generalize on bi-directional synthesis, for example, both A2G and G2A synthesis. Note that GCCDiff [16] also studies bi-directional CVIS. However, GCCDiff requires separate training for G2A and A2G synthesis, unlike our Geo^2 which only requires either A2G or G2A training to achieve bi-directional synthesis.

Geometric Foundation Models. Inspired by the success of foundation models in language and 2D vision [1, 4], Geometric Foundation Models (GFMs) have emerged as

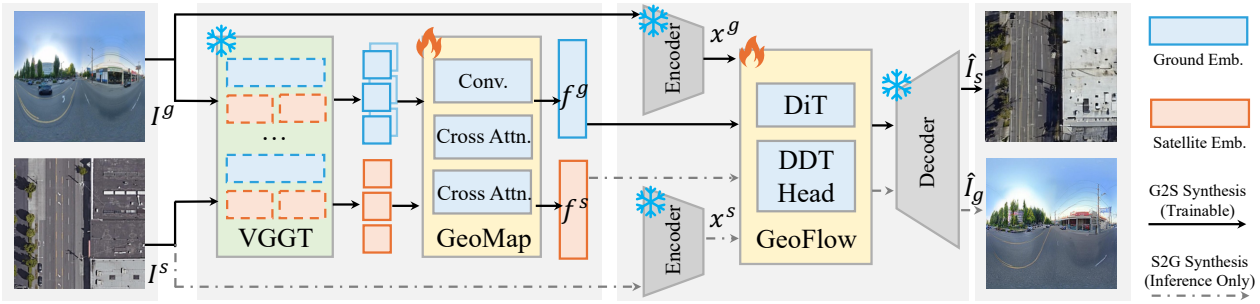

Figure 2. Overview of the Geo² framework. We first extract geometric features from ground and satellite images using VGGT. These dense features are then embedded into a shared geometry-aware latent space as detailed in Sec. 3.2. The resulting embeddings, f^g and f^s, are used for both CVGL and CVIS. While Geo² supports bidirectional image synthesis, it only requires training in the ground-to-satellite (G2S) direction. As detailed in Sec. 3.3, only ground images are needed as input during inference for G2S generation, and vice versa.

a promising solution for end-to-end 3D geometry prediction [6], which has wide application in downstream tasks like novel view synthesis [10, 52]. By pre-training on large-scale multi-view datasets, GFMs such as DUSt3R [37], MASt3R [14], and VGGT [35] are able to predict generalizable geometric attributes such as depth, point maps, and camera poses in a single forward pass. To improve reconstruction quality in more challenging scenarios like the aerial-ground domain, AerialMegaDepth [34] fine-tunes DUSt3R on aerial-ground imagery. GFMs have demonstrated the ability to reconstruct accurate geometric representations in multi-view settings [14, 34, 37]. Moreover, their dense features remain robust under sparse-view or even single-view conditions, where input images have little or no overlap [35]. However, how to leverage these robust geometric features for cross-view geo-spatial learning remains largely unexplored.

As shown in Table 1, most prior works treat CVGL and CVIS as separate tasks. In this paper, we propose Geo², a unified framework that jointly addresses both. To the best of our knowledge, we are the first to leverage the rich geometric representations from Geometric Foundation Models (GFMs) [14, 35, 37] for cross-view geo-spatial learning. As illustrated in Figure 1, directly feeding GFMs with cross-view image pairs often results in inaccurate geometric features due to large viewpoint differences and ground image distortions. To effectively extract usable geometry priors, we introduce GeoMap, a dual-branch model that processes ground and satellite images separately, as shown in Figure 3. GeoMap embeds the two views into a shared geometry-aware latent space, enforcing cross-view embedding alignment and improving CVGL. For CVIS, prior methods typically focus on single-direction synthesis (e.g., satellite-to-ground) [2, 15, 33, 45, 46]. These methods are based on GANs or diffusion models with strong assumptions, such as polar transformation [2, 33], which are not easily reversible. In contrast, we introduce flow matching [17] to model transformations between aerial and ground domains. This enables bi-directional synthesis from single-direction training, offering higher flexibility and data efficiency than existing methods.

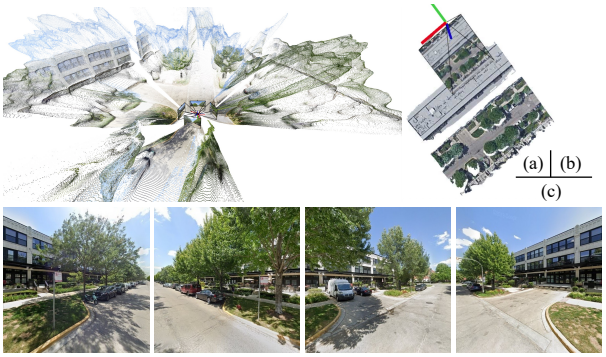

Figure 3. Illustration of VGGT reconstructions for (a) the ground view and (b) the satellite view, showing strong geometric alignment (e.g., buildings and overall layout). The ground view reconstruction is obtained from four perspective crops, illustrated in (c).

Note that prior efforts such as CDE [28] and RGCIS [44] also explored the combination of CVGL and CVIS. However, our approach differs fundamentally. CDE focuses on single-direction A2G synthesis, where a GAN and a SAFA [29] backbone are jointly trained. RGCIS relies on a frozen CVGL model to guide generation without mutual optimization. In contrast, Geo² jointly optimizes CVGL and bidirectional CVIS within a shared 3D-aware representation space, forming a coupled framework that unifies localization and generation for cross-view learning.

3. Methodology

We present an overview of Geo² in Figure 2. Section 3.1 provides a high-level description of the framework. In Section 3.2, we introduce GeoMap, which embeds both ground and satellite features into a shared latent space for improved geo-localization. Section 3.3 leverages these latent embeddings for direction image synthesis. Finally, Section 3.4 presents a joint training framework in which CVGL and CVIS mutually reinforce each other.

3.1. Geo² Overview

In this section, we first formulate the two cross-view geo-spatial tasks, CVGL and CVIS. We then present an

overview of the Geo^2 workflow.

Cross-view Geo-localization. This task focuses on retrieving a satellite image from a set of candidates given a ground-level image. Formally, the input consists of N ground-satellite image pairs $\{I_i^g, I_i^s\}_{i=1}^N$, where I^g and I^s denote ground and satellite images, respectively. Discriminative latent representations f_g and f_s are typically extracted from I^g and I^s by a localization model. Given a ground query image I_q^g with index q, the goal of CVGL is to retrieve the best matching satellite reference image I_b^s, where $q, b \in \{1, \ldots, N\}$, and a correct match is indicated when $b = q$. The objective can be formulated as: $b = \text{argmin}_{i \in \{1,\ldots,N\}} \|f_q^g - f_i^s\|_2$.

Cross-view Image Synthesis. This task focuses on generating a satellite view image $\hat{I}^s$ from a given ground-level image I^g. In this work, we address bidirectional image synthesis, where we also consider the reverse direction—generating a ground-level image $\hat{I}^g$ from a satellite image I^s. For simplicity, we refer to ground-to-satellite generation as G2S and satellite-to-ground as S2G. The objective of CVIS is to minimize the discrepancy between the generated image and the corresponding ground-truth image. This is commonly expressed as an L2 reconstruction loss: $\mathcal{L}_{\text{rec}} = \|\hat{I} - I\|_2$ where $\hat{I}$ and I represent the generated and reference images, respectively, in either direction.

Workflow. As illustrated in Figure 2, Geo^2 is a unified framework designed to jointly address the two tasks of CVGL and CVIS with geometric guidance from Geometric Foundation Models (GFMs). Given a pair of ground and satellite images, we first map them into a shared geometry-aware latent space using GeoMap. The resulting embeddings are directly used for the CVGL task, where retrieval is performed by computing the similarity between ground and satellite embeddings. The architecture of GeoMap and the embedding process are detailed in Section 3.2.

Importantly, the shared latent space also facilitates bidirectional image synthesis. To exploit this, we introduce a flow-matching model, described in Section 3.3, which is conditioned on the geometry-aware latent embeddings. The flow model is reversible and supports both ground-to-satellite and satellite-to-ground generation. Finally, in Section 3.4, we present a joint training scheme where CVGL and CVIS are optimized together. This allows the two tasks to mutually reinforce each other during fine-tuning, further improving overall performance.

3.2. Geo-localization with GeoMap

While GFMs provide strong geometric priors, it remains largely unexplored how such information can benefit cross-view geo-spatial tasks like CVGL. Directly feeding ground-satellite image pairs into GFMs often fails to produce geometry-consistent features due to the large viewpoint gap between ground and aerial imagery [34]. To address

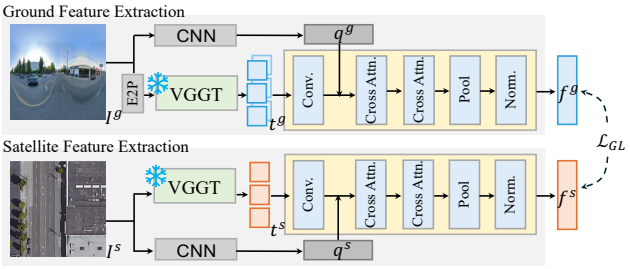

Figure 4. Overview of GeoMap pipeline. Ground and satellite images are individually processed via two separate branches.

this, we propose GeoMap, a dual-branch model that maps ground and satellite images into a shared latent space, effectively integrating geometric priors from GFMs.

Geometric Feature Extraction. Given ground images I^g and satellite images I^s, we first extract geometry-aware features t^g and t^s using VGGT [35]. Since satellite images can be treated as single-perspective inputs, they can be directly processed by VGGT, as $t^s = \text{VGGT}(I^s)$, where $t^s \in \mathbb{R}^{C \times H_1 \times W_1}$, and C is the feature dimension. While VGGT is capable of handling multi-view input images, directly applying it to ground images is challenging. This is because, in the CVGL task, ground images are often panoramas represented in equiangular format and lie in spherical coordinates, which introduce significant camera distortions. These distortions degrade the quality of the extracted features, as VGGT is primarily trained on perspective imagery. To this end, we apply an equiangular-to-perspective (E2P) transformation to convert the ground image into multiple perspective crops, as shown in Figure 3,

$$\{IP^i\}_{i=1}^V = \text{E2P}(I^g), \tag{1}$$

which results in a set of V perspective views that densely cover the horizontal field of view. Further details on the coordinate transformation from spherical to perspective are provided in the Supplementary. These V perspective views are then fed into VGGT in a multi-view inference setup to produce the feature $t^g \in \mathbb{R}^{V \times C \times H_1 \times W_1}$. Finally, we embed the geometry features t^g and t^s into a shared latent space.

Feature Embedding. The retrieval in CVGL is typically performed using embeddings of dimension D, which requires aggregating information from the high-dimensional features t^g and t^s. To achieve this, we extract feature maps with the target dimension D by applying a convolutional layer to the VGGT features. Specifically, we obtain $t^{s\prime} = \text{Conv}(t^s)$, where $t^{s\prime} \in \mathbb{R}^{D \times H_1' \times W_1'}$. Similarly, for the ground image, we get $t^{g\prime} \in \mathbb{R}^{V \times D \times H_1' \times W_1'}$, where V is the number of perspective crops and remains unchanged.

In parallel, a pretrained CNN is used to extract semantic features from the original satellite and ground images. These features are flattened into token sequences $q^s \in \mathbb{R}^{D \times N_1}$ and $q^g \in \mathbb{R}^{D \times N_2}$, respectively. We also flatten $t^{s\prime}$ and $t^{g\prime}$ along the spatial dimensions to obtain dense

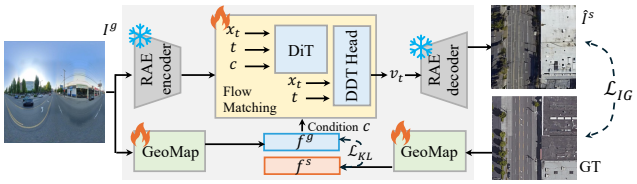

Figure 5. Overview of our GeoFlow pipeline. The latent representation (f^g if ground-to-satellite synthesis or f_s if satellite-to-ground synthesis) is input as condition C.

tokens of the same dimension D. The semantic tokens q^s and q^g are treated as query tokens and aggregate information from $t^{s\prime}$ and $t^{g\prime}$ via cross-attention. For the satellite branch, this is formulated as,

$$out^s = \text{Attn}(q^s, t^{s\prime}, t^{s\prime}). \tag{2}$$

The final satellite embedding f^s is obtained by applying average pooling followed by normalization on out^s. The ground embedding f^g is computed in the same manner. These representations, f^s and f^g, lie in a shared geometry-aware latent space and encode both semantic and geometric information, which are used as the final embeddings for cross-view geo-localization. Optimization is guided by the InfoNCE loss [24, 27] $\mathcal{L}_{GL}$, with training details described in Section 3.4.

3.3. Cross-view Image Synthesis with GeoFlow

The target of Cross-View Image Synthesis (CVIS) is to generate satellite images from given corresponding ground images, or vice versa. Existing methods typically frame the problem as a conditional generation problem [15, 21, 31, 33, 45]. We argue that CVIS is better framed as a domain translation problem by leveraging the flow matching [17] framework, leading to a more versatile training and inference pipeline. The overview of our GeoFlow pipeline is presented in Figure 5.

Single Directional Training. Consider a pair of satellite and ground image I^g and I^s, we took the pretrained RAE [53] to encode the image into latent space, denoting as x^g and x^s, respectively. We define the probability path by applying the optimal transport displacement interpolation [19], such that,

$$x_t = (1-t) \times x^g + t \times x^s, \tag{3}$$

where $t \in [0, 1]$. Following the principle of flow matching [17], we train a Network G_θ to predict the vector field $v = x^s - x^g$. Thus, the loss is defined as,

$$\mathcal{L}_{IG} = \|G_\theta(x_t, t, c) - v\|_2, \tag{4}$$

where c is the learned embedding from our GeoMap (Section 3.1) as an auxiliary condition. We employ lightweight DiTs [25] with DDT heads [38] as the backbone of G_θ. To better align the auxiliary condition c, we use a consistency loss $\mathcal{L}_{KL}$ on f_g and f_s, detailed in Section 3.4.

Bi-Directional Synthesis: An advantage of our Geo^2 is to achieve bi-directional synthesis without retraining. Mathematically, the trained G_θ defines an ODE function. To solve this equation, we can apply the following integral,

$$x^s = x^g + \int_0^1 G_\theta(x_t, t, c)\, dt. \tag{5}$$

By simply reversing the direction of the integral,

$$\begin{aligned} x^g &= x^s + \int_1^0 G_\theta(x_t, t, c)\, dt \\ &= x^s - \int_0^1 G_\theta(x_t, t, c)\, dt, \end{aligned} \tag{6}$$

in this way, we can generate ground images from satellite images through Equation (6), even if the model has never been trained in this direction.

3.4. Joint Training

As mentioned in Section 3.2 and Section 3.3, both GeoMap and GeoFlow leverages the embeddings f_s and f_g in the shared geometry-aware latent space. In this section, we present our joint training strategy where CVGL and CVIS can mutually benefit each other, as shown in Algorithm 1.

Algorithm 1: Joint Cross-View Training

Input : Ground-Satellite Image Pairs, $\{I^g, I^s\}$
Output : GeoMap model M_β and GeoFlow model G_θ.
Initialize model parameters β and θ.
for *epoch* $= 1$ **to** T_1 **do**
  $f^g, f^s \leftarrow M_\beta(I^g, I^s)$ `// Extract embeddings`
  Compute $\mathcal{L}_{GL}$ `// InfoNCE loss`
  $\beta \leftarrow \beta - \eta_1 \nabla_\beta \mathcal{L}_{GL}$ `// Update model parameters`
for *epoch* $= 1$ **to** T_3 **do**
  $c \leftarrow f^g; f^g, f^s \leftarrow M_\beta(I^g, I^s)$ `// RAE omitted`
  $x^s \leftarrow x^g + \int_0^1 G_\theta(x_t, t, c)\, dt$ `// Generate image`
  Compute $\mathcal{L}_{IG}$ `// Generation loss`
  $\theta \leftarrow \theta - \eta_2 \nabla_\theta \mathcal{L}_{GL}$
  if *not epoch* $< T_2$ **then**
    Compute $\mathcal{L}_{KL}$ `// Consistency loss`
    $\beta \leftarrow \beta - \eta_3 \nabla_\beta (\mathcal{L}_{GL} + \alpha \mathcal{L}_{KL})$
return M_β and G_θ

We use M_β to denote GeoMap model parameterized by β, and G_θ for GeoFlow model parameterized by θ. We use both task-specific and joint objectives to optimize the models. Specifically, in the first stage, we froze the CNN and VGGT backbones in GeoMap, and use $\mathcal{L}_{GL}$ to optimize only M_β for T_1 epochs. The trained M_β can embed ground and satellite images in a geometry-aware shared latent space. In stage two, we first train the GeoFlow model G_θ for T_2 epochs. Then, with both trained GeoMap and GeoFlow models, we jointly fine-tune M_β and G_θ for $T_3 - T_2$

epochs with an additional consistency loss $\mathcal{L}_{KL}$. We provide more details on the losses below.

Task Specific Optimization. The CVGL task is often optimized via the contrastive loss $\mathcal{L}_{GL}$, such as triplet loss [50, 51] and InfoNCE loss [8, 24, 27]. We train our GeoMap with the InfoNCE loss. Given a ground embeddings f^g and a batch of N satellite embeddings $\{f_i^s\}_{i=1}^N$, in which only f_+^s matches the ground query, we have,

$$\mathcal{L}_{GL} = -\log \frac{\exp\left(f^g \cdot f_+^s/\tau\right)}{\sum_{i=1}^{N} \exp\left(f^g \cdot f_i^s/\tau\right)}, \tag{7}$$

where τ is the hyperparameter controlling the softness of the distribution. Under $\mathcal{L}_{GL}$, a positive example f_+^s is effectively contrasted by $N-1$ negative examples in the batch, which optimize the similarity of matching pairs. For CVIS, we use the L_2 loss described in Section 3.3.

Joint Optimization. Since the geometry-aware shared latent space of ground and satellite embeddings benefits both CVGL and CVIS, it is natural to jointly optimize GeoMap and GeoFlow to enhance the consistency of this latent space. Therefore, after task specific training, we fine-tune GeoMap and GeoFlow together with a consistency loss,

$$\mathcal{L}_{KL} = \mathrm{KL}(f^g \,\|\, f^s) + \mathrm{KL}(f^s \,\|\, f^g), \tag{8}$$

where we align the distribution of f^g and f^s more explicitly. We empirically find the consistency loss improve both retrieval accuracy and bidirectional generation quality. More results can be found in Section 4.

4. Experiments

Datasets and Baselines: To demonstrate the effectiveness of Geo^2, we benchmark on three popular datasets, CVUSA [41], CVACT [18], and VIGOR [54] on both CVIGL and CVIS tasks. **CVUSA** [41] contains 35,532 training panoramic ground and satellite image pairs and 8,884 testing pairs, sparsely sampled from the United States of America. **CVACT** [18] provides similar number of training and validation (CVACT val) pairs as the CVUSA dataset. Moreover, it provides a challenging testing set (CVACT test), which contains 92,802 ground and satellite pairs. The training, validation, and testing sets of CVACT are densely collected from Canberra, Australia, and geographically split. **VIGOR** [54] incorporates a many-to-one configuration which collects ground panoramas and satellite images from 4 cities in the U.S, namely, New York, Chicago, Seattle, and San Francisco, resulting in 90,618 ground panoramas and 105,124 satellite images. Furthermore, VIGOR provides the same-area evaluation and the cross-area evaluation. In which the same-area protocol stands for training and testing on all 4 cities, and the cross-area protocol means training on 2 cities (New York and Seattle) and testing on the other 2 cities (Chicago and San Francisco). *For more implementation details and parameter settings, please refer to our supplementary material.*

Table 2. Comparison of cross-view geo-localization performance on the VIGOR dataset under same-area and cross-area settings. We report recall rates (%) and hit rate (%) at different top-K retrieval thresholds. The best results are shown in **bold** and the second-best results are underlined.

Dataset	Approach	R@1	R@5	R@10	R@1%	Hit Rate
VIGOR Same-area	SAFA† [29]	33.93	58.42	68.12	98.24	36.87
	TransGeo [55]	61.48	87.54	91.88	99.56	73.09
	SAIG-D [56]	65.23	88.08	-	99.68	74.11
	GeoDTR [50]	56.51	80.37	86.21	99.25	61.76
	GeoDTR+ [51]	59.01	81.77	87.10	99.07	67.41
	Sample4Geo [8]	77.86	95.66	97.21	99.61	89.82
	PanoBEV [46]	**82.18**	**97.10**	98.17	**99.70**	-
	Ours	81.59	96.53	**98.62**	99.68	**90.35**
VIGOR Cross-area	SAFA† [29]	8.20	19.59	26.36	77.61	8.85
	TransGeo [55]	18.99	38.24	46.91	88.94	21.21
	SAIG-D [56]	33.05	55.94	-	94.64	36.71
	GeoDTR [50]	30.02	52.67	61.45	94.40	30.19
	GeoDTR+ [51]	36.01	59.06	67.22	94.95	39.40
	Sample4Geo [8]	61.70	83.50	88.00	98.17	69.87
	PanoBEV [46]	**72.19**	**88.68**	**91.68**	**98.56**	-
	Ours	66.71	87.34	91.02	98.25	**72.13**

Evaluation Metrics: To evaluate Geo^2, we independently benchmark its performance on CVIS and CVIGL tasks. By following existing CVIS works [15, 16, 26, 48], we adopt Structure Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), LPIPS [49], and FID score [9]. For the CVIGL task, we adopt the Recall accuracy at Top-K (R@K), which evaluates if the ground truth of the query ground image is measured at the top K retrieved results from the satellite reference set. Conventionally, K is chosen to be 1, 5, 10, and 1%. For the VIGOR dataset, we also evaluate the hit rate, which measures whether the top-1 retrieved satellite image covers the query ground image location.

4.1. Quantitative Comparison on CVGL

Quantitative comparisons between our Geo^2 and existing state-of-the-art on three cross-view geo-localization benchmarks are summarized in Table 3 and Table 2. While performance on CVUSA [41] is already near saturation, our method still pushes the boundary further. However, the advantages of Geo^2 are most pronounced on the more difficult benchmarks. On the CVACT validation set and the CVACT testing set, Geo^2 demonstrates a significant performance gain, 2.46% on CVACT Val and 1.40% on CVACT Test on R@1 accuracy. Furthermore, on the challenging VIGOR [54] dataset, comparing with the well-established Sample4Geo [8] baseline, we improve R@1 by 3.73% in the Same-Area setting and by a significant 5.01% in the challenging Cross-Area setting. These results validate our hypothesis: incorporating geometric foundation models provides critical 3D spatial priors, resulting in feature representations that remain robust even when the visual ap-

Table 3. Comparison of cross-view geo-localization performance on CVUSA and CVACT datasets in recall at top-K retrieves (R@K). The best results are shown in **bold** and the second-best results are underlined. † indicates Polar Transformation is applied.

Approach	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%
LPN [39]	85.79	95.38	96.98	99.41	79.99	90.63	92.56	-	-	-	-	-
SAFA† [29]	89.84	96.93	98.14	99.64	81.03	92.80	94.84	-	-	-	-	-
TransGeo [55]	94.08	98.36	99.04	99.77	84.95	94.14	95.78	98.37	-	-	-	-
GeoDTR [50]	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† [50]	95.43	98.86	99.34	99.86	86.21	95.44	96.72	98.77	64.52	88.59	91.96	98.74
SAIG-D† [56]	96.34	99.10	99.50	99.86	89.06	96.11	97.08	98.89	67.49	89.39	92.30	96.80
Sample4Geo [8]	98.68	99.68	99.78	99.87	90.35	96.61	97.53	98.78	71.51	92.42	94.45	98.70
PanoBEV [46]	98.71	99.70	99.78	99.86	91.90	97.23	97.84	98.84	73.68	93.53	95.11	98.81
Ours	**98.83**	**99.72**	**99.79**	**99.91**	**94.36**	**97.41**	**97.97**	**99.05**	**75.08**	**94.89**	**95.77**	**99.01**

pearance changes drastically between training and testing. To further benchmark the generalization of Geo^2, we experiment with it on the cross-dataset test, which includes two sub-tasks: 1) training on CVUSA [41] and testing on CVACT [18] (CVUSA $\rightarrow$ CVACT), and vice versa (CVACT $\rightarrow$ CVUSA). As shown in Table 4, Geo^2 illustrates a strong generalization capability in the cross-dataset test. Notably, on CVACT $\rightarrow$ CVUSA, Geo^2 achieves a score of 55.14% on R@1, leading a 10.19% leap from Sample4Geo [8]. On CVUSA $\rightarrow$ CVACT, Geo^2 also achieves 63.17%, which is comparable with the existing state-of-the-art. To conclude, the quantitative experiments on all three datasets show the superior performance of our Geo^2 on same-area, cross-area, and cross-dataset evaluations. The results demonstrate the advantage of the geometric foundation models, not only improving the discrimination of the feature representations but also enhancing the generalization of Geo^2 on unseen data.

4.2. Quantitative Comparison on CVIS

As discussed in Section 3.3, our Geo^2 cannot only achieve cross-view geo-localization but also can perform Cross-View Image Synthesis (CVIS). In this section, we evaluate the CVIS performance of Geo^2 on CVUSA [41], CVACT [18], and VIGOR [54] datasets. Since the uniqueness of our Geo^2 that can perform bi-directional synthesis without **re-training**, we conduct evaluation on both Ground-to-Satellite (G2S) and Satellite-to-Ground (S2G). The results are summarized in Table 5 and Table 6, respectively. Note that Geo^2 can perform significantly better than existing methods on the CVACT dataset on all the evaluation metrics, especially FID scores, which decrease from 36.48 to 31.72. Geo^2 also achieves outstanding performance on the other two datasets. For example, Geo^2 has a FID score of 30.09 on the VIGOR dataset, while bringing the LPIPS down to 0.594. On the S2G task, Geo^2 is not as good as it is on the G2S task. However, it still achieves the best FID score on the CVACT and CVUSA dataset, while keeping the other scores comparable to the baseline methods. In summary, the evaluation results reveal the superiority of Geo^2 on the cross-view image synthesis task. It also demonstrates the flexibility of Geo^2 that can perform bi-directional synthesis without re-training. We attribute the performance improvement to the introduction of geometric foundation models that inject 3D priors into the flow matching process and also the rich latent representation from the RAE [53] encoder and decoder. *For more experiments and analysis, please refer to our supplementary material.*

Table 4. Comparison of cross-view geo-localization performance on cross-dataset benchmarks. CVUSA $\rightarrow$ CVACT stands for training on CVUSA and testing on CVACT. CVACT $\rightarrow$ CVUSA stands for training on CVACT and testing on CVUSA. The best results are shown in **bold** and the second-best results are underlined.

Dataset	Approach	R@1	R@5	R@10	R@1%
CVUSA ↓ CVACT	SAFA [29]	30.40	52.93	62.29	85.82
	TransGeo [55]	37.81	61.57	69.86	89.14
	SAIG-D [56]	15.29	33.07	42.14	72.95
	GeoDTR [50]	43.72	66.99	74.61	91.83
	GeoDTR+ [51]	60.16	79.97	84.67	94.48
	Sample4Geo [8]	56.62	77.79	87.02	94.69
	PanoBEV [46]	**67.79**	**84.06**	**87.96**	95.05
	Ours	63.17	82.53	87.88	**95.09**
CVACT ↓ CVUSA	SAFA [29]	21.45	36.55	43.79	69.83
	TransGeo [55]	17.45	32.49	40.48	69.14
	SAIG-D [56]	18.97	35.60	44.28	75.33
	GeoDTR [50]	29.85	49.25	57.11	82.47
	GeoDTR+ [51]	52.56	73.08	79.82	94.80
	Sample4Geo [8]	44.95	64.36	72.10	90.65
	PanoBEV [46]	44.10	70.68	75.86	95.31
	Ours	**55.14**	**73.58**	**80.03**	**95.33**

4.3. Qualitative Visualization

We visualize synthesized results on CVUSA [41], CVACT [18], and VIGOR [54], in Figure 6, Figure 7, and Figure 8, respectively. We randomly select 3 samples for each dataset and include diverse scenarios, such as urban, suburban, and rural areas. For each sample, we present both Ground-to-Satellite (G2S) and Satellite-to-Ground (S2G) synthesized images. As we can see in Figure 6, our Geo^2 can synthesize geometrically-matching and high quality satellite images and ground images. Specifi-

Table 5. Comparison of Ground-to-Satellite image synthesis performance on CVUSA, CVACT, and VIGOR datasets. We report FID [9], LPIPS [49], PSNR and SSIM scores. The best results are shown in **bold**.

Approach	CVUSA				CVACT				VIGOR			
	FID (↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	FID (↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	FID (↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)
X-Seq [28]	161.16	0.706	11.97	0.084	190.12	0.661	12.41	0.042	-	-	-	-
Aerial Diff [13]	136.18	0.855	10.06	0.103	127.29	0.878	10.24	0.108	123.16	0.831	11.49	0.141
GPG2A [2]	58.80	0.691	12.13	0.135	63.50	0.690	11.98	0.116	70.19	0.695	**11.81**	0.159
ControlNet [48]	32.45	0.650	12.63	0.149	62.21	0.682	11.95	0.115	53.27	0.666	10.38	0.170
Skydiffusion [45]	**29.18**	0.635	**14.58**	**0.168**	36.48	0.645	12.85	0.118	45.29	0.661	11.69	**0.186**
Ours	33.68	**0.534**	13.83	0.167	**31.72**	**0.552**	**14.62**	**0.162**	**30.09**	**0.594**	11.33	0.127

Table 6. Comparison of Satellite-to-Ground image synthesis performance on CVUSA, CVACT, and VIGOR datasets. We report FID [9], LPIPS [49], PSNR and SSIM scores. The best results are shown in **bold**.

Approach	CVUSA				CVACT				VIGOR			
	FID (↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	FID (↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	FID (↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)
Sat2Density [26]	41.43	**0.4163**	**14.66**	**0.358**	47.09	**0.3339**	**16.38**	**0.482**	47.98	0.3488	11.21	0.229
ControlNet [48]	51.48	0.5158	10.91	0.148	49.41	0.4563	11.66	0.229	53.29	0.4051	10.59	0.191
CrossViewDiff [15]	**23.67**	0.4412	12.00	0.371	41.94	0.3661	12.41	0.412	26.57	**0.3414**	**13.68**	0.268
Ours	29.05	0.4871	12.94	0.317	**27.77**	0.4833	13.57	0.457	**22.90**	0.4690	12.64	**0.380**

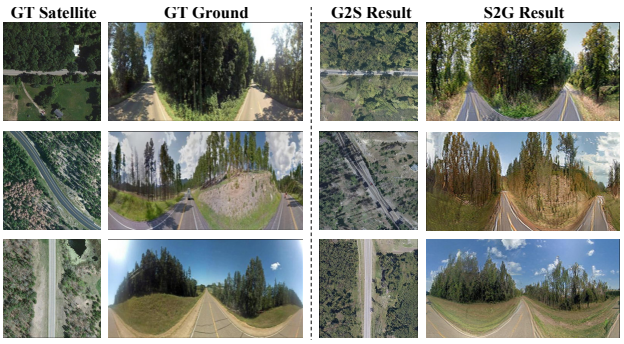

Figure 6. Visualization of Generated images from Geo^2 on CVUSA dataset. From left to right are the ground truth satellite image, the ground truth ground image, the Ground-to-Satellite generated image, and the Satellite-to-Ground generated image.

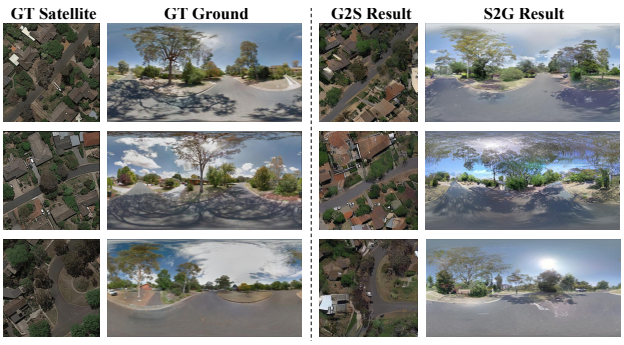

Figure 7. Visualization of Generated images from Geo^2 on CVACT dataset. The order is the same as Figure 6.

cally, by comparing the first and third example in Figure 6, our model can accurately capture the road orientation and the positions of the side objects, such as trees and grassland. Figure 7 illustrates more complex scenarios in urban and suburban areas, for example, in the third row, our Geo^2 successfully captures the curvature of the road and is reconstructed in the synthesized satellite image. The most challenging cases arise from the VIGOR dataset, which includes

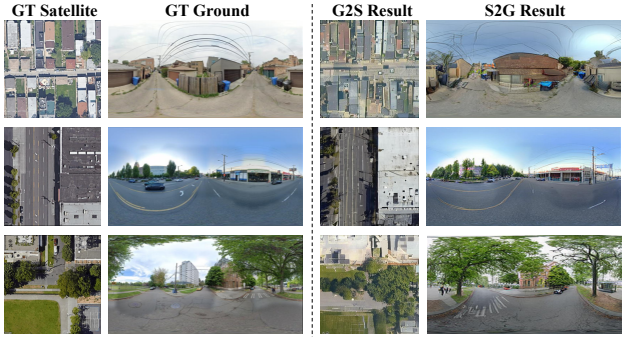

Figure 8. Visualization of Generated images from Geo^2 on VIGOR dataset. The order is the same as Figure 6.

complex scenarios. Surprisingly, our model also captures both the high-level structural information and low-level details. For instance, in the first example, Geo^2 reconstructs the correct building layout in both G2S and S2G results. In the second example, Geo^2 successfully models the store on the right side of the road and the trees on the left side of the road. In summary, the qualitative visualization illustrates the outstanding image synthesis quality. *For more qualitative results, please refer to our supplementary material.*

5. Conclusion

In this paper, we present Geo^2, a geometry-guided framework that unifies cross-view geo-spatial tasks. We first propose GeoMap, a dual-branch model that embeds ground and satellite images into a shared geometry-aware latent space. These embeddings are directly used for CVGL and further serve as conditioning inputs for bidirectional CVIS in our GeoFlow module. We introduce a joint training strategy, allowing the two tasks to mutually benefit. Geo^2 explores 3D geometric priors from GFMs as a new and effective means of bridging CVGL and CVIS, offering a promising direction for unified cross-view geo-spatial learning.

Acknowledgments: This work was supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number 140D0423C0074. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes, notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI/IBC, or the U.S. Government.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023. 2

[2] Ahmad Arrabi, Xiaohan Zhang, Waqas Sultani, Chen Chen, and Safwan Wshah. Cross-view meets diffusion: Aerial image synthesis with geometry and text guidance. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5356–5366. IEEE, 2025. 1, 2, 3, 8, 5

[3] Sudong Cai, Yulan Guo, Salman Khan, Jiwei Hu, and Gongjian Wen. Ground-to-aerial image geo-localization with a hard exemplar reweighting triplet loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2

[4] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything, 2023. 2

[5] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training. *arXiv preprint arXiv:2509.26645*, 2025. 3

[6] Wenyan Cong, Yiqing Liang, Yancheng Zhang, Ziyi Yang, Yan Wang, Boris Ivanovic, Marco Pavone, Chen Chen, Zhangyang Wang, and Zhiwen Fan. E3d-bench: A benchmark for end-to-end 3d geometric foundation models. *arXiv preprint arXiv:2506.01933*, 2025. 3

[7] Quan Dao, Hao Phung, Trung Tuan Dao, Dimitris N Metaxas, and Anh Tran. Self-corrected flow distillation for consistent one-step and few-step image generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2654–2662, 2025. 6

[8] Fabian Deuser, Konrad Habel, and Norbert Oswald. Sample4geo: Hard negative sampling for cross-view geo-localisation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16847–16856, 2023. 2, 6, 7, 1, 3, 5

[9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2017. 6, 8

[10] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139:1–139:14, 2023. 3

[11] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 3

[12] Nikita Kornilov, Petr Mokrov, Alexander Gasnikov, and Alexander Korotin. Optimal flow matching: Learning straight trajectories in just one step. *Advances in Neural Information Processing Systems*, 37:104180–104204, 2024. 6

[13] Divya Kothandaraman, Tianyi Zhou, Ming Lin, and Dinesh Manocha. Aerial diffusion: Text guided ground-to-aerial view synthesis from a single image using diffusion models. In *SIGGRAPH Asia 2023 Technical Communications*, pages 1–4. 2023. 8, 2

[14] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 1, 3

[15] Weijia Li, Jun He, Junyan Ye, Huaping Zhong, Zhimeng Zheng, Zilong Huang, Dahua Lin, and Conghui He. Crossviewdiff: A cross-view diffusion model for satellite-to-street view synthesis. *arXiv preprint arXiv:2408.14765*, 2024. 2, 3, 5, 6, 8

[16] Tao Jun Lin, Wenqing Wang, Yujiao Shi, Akhil Perincherry, Ankit Vora, and Hongdong Li. Geometry-guided cross-view diffusion for one-to-many cross-view image synthesis. *arXiv preprint arXiv:2412.03315*, 2024. 2, 6

[17] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022. 3, 5

[18] Liu Liu and Hongdong Li. Lending orientation to neural networks for cross-view geo-localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5624–5633, 2019. 2, 6, 7

[19] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.03003*, 2022. 5

[20] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11976–11986, 2022. 1, 3

[21] Xiaohu Lu, Zuoyue Li, Zhaopeng Cui, Martin R Oswald, Marc Pollefeys, and Rongjun Qin. Geometry-aware satellite-to-ground image synthesis for urban areas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 859–867, 2020. 1, 2, 5

[22] Zonglin Lyu and Chen Chen. Tlb-vfi: Temporal-aware latent brownian bridge diffusion for video frame interpolation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 4

[23] Zonglin Lyu, Ming Li, Jianbo Jiao, and Chen Chen. Frame interpolation with consecutive brownian bridge diffusion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024. 4

[24] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5, 6

[25] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4195–4205, 2023. 5, 2

[26] Ming Qian, Jincheng Xiong, Gui-Song Xia, and Nan Xue. Sat2density: Faithful density learning from satellite-ground image pairs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3683–3692, 2023. 1, 2, 6, 8, 5

[27] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 5, 6

[28] Krishna Regmi and Ali Borji. Cross-view image synthesis using conditional gans. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 3501–3510, 2018. 2, 3, 8, 5

[29] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geo-localization. *Advances in Neural Information Processing Systems*, 32:10090–10100, 2019. 1, 2, 3, 6, 7

[30] Yujiao Shi, Xin Yu, Dylan Campbell, and Hongdong Li. Where am i looking at? joint location and orientation estimation by cross-view matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[31] Yujiao Shi, Dylan Campbell, Xin Yu, and Hongdong Li. Geometry-guided street-view panorama synthesis from satellite imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10009–10022, 2022. 2, 5

[32] Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2417–2426, 2019. 2

[33] Aysim Toker, Qunjie Zhou, Maxim Maximov, and Laura Leal-Taixé. Coming down to earth: Satellite-to-street view synthesis for geo-localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6488–6497, 2021. 2, 3, 5

[34] Khiem Vuong, Anurag Ghosh, Deva Ramanan, Srinivasa Narasimhan, and Shubham Tulsiani. Aerialmegadepth: Learning aerial-ground reconstruction and view synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21674–21684, 2025. 1, 3, 4

[35] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 2, 3, 4, 6

[36] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 3

[37] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 1, 3

[38] Shuai Wang, Zhi Tian, Weilin Huang, and Limin Wang. Ddt: Decoupled diffusion transformer. *arXiv preprint arXiv:2504.05741*, 2025. 5, 2

[39] 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*, pages 1–1, 2021. 7, 2

[40] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Image and object geo-localization. *International Journal of Computer Vision*, 132(4):1350–1392, 2024. 2

[41] Scott Workman, Richard Souvenir, and Nathan Jacobs. Wide-area image geolocalization with aerial reference imagery. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3961–3969, 2015. 2, 6, 7

[42] Songsong Wu, Hao Tang, Xiao-Yuan Jing, Haifeng Zhao, Jianjun Qian, Nicu Sebe, and Yan Yan. Cross-view panorama image synthesis. *IEEE Transactions on Multimedia*, 2022. 2

[43] Hongji Yang, Xiufan Lu, and Yingying Zhu. Cross-view geo-localization with layer-to-layer transformer. In *Advances in Neural Information Processing Systems*, pages 29009–29020. Curran Associates, Inc., 2021. 2

[44] Hongji Yang, Yiru Li, and Yingying Zhu. Retrieval-guided cross-view image synthesis. *arXiv preprint arXiv:2411.19510*, 2024. 2, 3

[45] Junyan Ye, Jun He, Weijia Li, Zhutao Lv, Jinhua Yu, Haote Yang, and Conghui He. Skydiffusion: Street-to-satellite image synthesis with diffusion models and bev paradigm. *arXiv preprint arXiv:2408.01812*, 2024. 1, 2, 3, 5, 8

[46] 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. In *European Conference on Computer Vision*, pages 74–90. Springer, 2024. 2, 3, 6, 7

[47] Menghua Zhai, Zachary Bessinger, Scott Workman, and Nathan Jacobs. Predicting ground-level scene layout from aerial imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[48] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 6, 8, 2

[49] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6, 8, 4

[50] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Yi Zhou, and Safwan Wshah. Cross-view geo-localization via learning

disentangled geometric layout correspondence. In *Proceedings of the AAAI conference on artificial intelligence*, pages 3480–3488, 2023. 1, 2, 6, 7

[51] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: toward generic cross-view geolocalization via geometric disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 6, 7

[52] Yancheng Zhang, Guangyu Sun, and Chen Chen. EGGS: Exchangeable 2d/3d gaussian splatting for geometry-appearance balanced novel view synthesis. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 3

[53] Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. *arXiv preprint arXiv:2510.11690*, 2025. 5, 7, 2, 3

[54] Sijie Zhu, Taojiannan Yang, and Chen Chen. Vigor: Cross-view image geo-localization beyond one-to-one retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3640–3649, 2021. 2, 6, 7

[55] Sijie Zhu, Mubarak Shah, and Chen Chen. 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)*, pages 1162–1171, 2022. 2, 6, 7, 3

[56] Yingying Zhu, Hongji Yang, Yuxin Lu, and Qiang Huang. Simple, effective and general: A new backbone for cross-view image geo-localization. *arXiv preprint arXiv:2302.01572*, 2023. 2, 6, 7