Cross-View Meets Diffusion: Aerial Image Synthesis With Geometry And Text Guidance

Ahmad Arrabi[1†], Xiaohan Zhang[1†], Waqas Sultani[2], Chen Chen[3], Safwan Wshah[1*]
[1] Vermont Artificial Intelligence Lab, Department of Computer Science, University of Vermont
[2] Intelligent Machines Lab, Information Technology University
[3]Center for Research in Computer Vision, University of Central Florida
[†] These authors contributed equally. [*] Corresponding and senior author.

Abstract

*Aerial imagery analysis is critical for many research fields. However, obtaining frequent high-quality aerial images is not always accessible due to its high effort and cost requirements. One solution is to use the Ground-to-Aerial (G2A) technique to synthesize aerial images from easily collectible ground images. However, G2A is rarely studied, because of its challenges, including but not limited to, the drastic view changes, occlusion, and range of visibility. In this paper, we present a novel **G**eometric **P**reserving Ground-to-Aerial (**G2A**) image synthesis (GPG2A) model that can generate realistic aerial images from ground images. GPG2A consists of two stages. The first stage predicts the Bird's Eye View (BEV) segmentation (referred to as the BEV layout map) from the ground image. The second stage synthesizes the aerial image from the predicted BEV layout map and text descriptions of the ground image. To train our model, we present a new multi-modal cross-view dataset, namely VIGORv2, built upon VIGOR [64] with newly collected aerial images, maps, and text descriptions. Our extensive experiments illustrate that GPG2A synthesizes better geometry-preserved aerial images than existing models. We also present two applications, data augmentation for cross-view geo-localization and sketch-based region search, to further verify the effectiveness of our GPG2A. The code and dataset are available at* https://github.com/AhmadArrabi/GPG2A.

1. Introduction

Unlike satellite images, which are low in resolution and can be obscured by clouds [6, 11, 20], aerial images capture more detailed views, benefiting various applications, such as land use classification [4, 52], urban planning [45], transportation [10, 21, 29], socioeconomic studies [5, 32], and cross-view geo-localization (CVGL) [41, 50, 59, 60, 64].

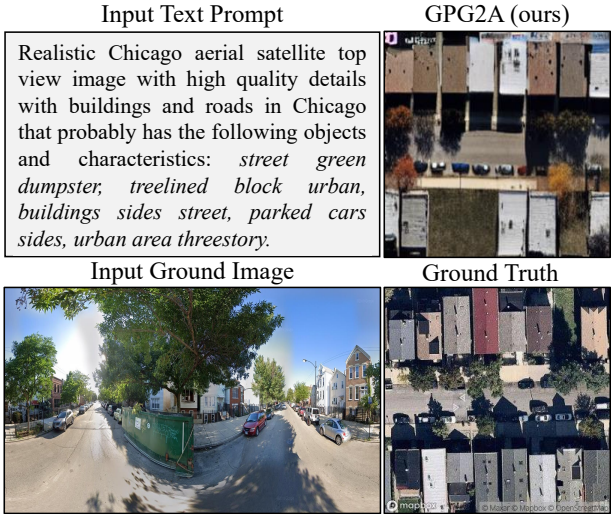

Figure 1. An example generated aerial image (top right) by our GPG2A from the input text prompt (top left) and the ground image (bottom left). The ground truth aerial image is on the bottom right.

However, current aerial images are limited by the high effort and cost required to capture them, as they are often captured by Unmanned Aerial Vehicles (UAVs) or drones. For example, New York State's government annually captures aerial images for only one-third of its counties [3]. Security concerns also restrict drone use at low altitudes in urban areas, limiting applications and preventing frequent updates. These accessibility challenges are more common in developing countries. In contrast, ground images are far more available and cost-effective, especially in the recent advanced cars and autonomous vehicles. Also, crowdsourcing platforms like Mapillary [2] see tons of daily uploads of street-view images. *Thus, a promising solution for such challenges is ground-to-aerial (G2A) image synthesis*, which aims to generate more frequent aerial images from their corresponding ground views.

Despite the potential of G2A image synthesis, to the best of our knowledge, there has been limited research address-

ing this task due to its challenges. These challenges include the drastic viewing angle change, object occlusions, and different ranges of visibility between aerial and ground views. Some prior works attempted G2A synthesis mainly leveraging Generative Adversarial Networks (GANs) [15] but lacked explicit geometric constraints [31] or depended on strong priors like segmentation maps of the aerial view [44].

In this work, we propose **G**eometric **P**reserving Ground-to-Aerial (**G2A**) image synthesis (GPG2A) model which features a novel two-stage process. The first stage transforms the input ground image into a Bird's Eye View (BEV) layout map. The second stage leverages pre-trained diffusion models [27, 57], conditioned on the predicted BEV layout map from the first stage, to generate photo-realistic aerial images This innovative two-stage pipeline provides three advantages: 1) The problem is simplified by introducing an intermediate BEV layout map stage reducing the domain gap between aerial and ground views. 2) The BEV layout map explicitly preserves the geometry, enhancing the synthesized aerial images by maintaining consistent geometry with ground images and reducing overfitting to low-level details. 3) By leveraging the pre-trained knowledge from diffusion foundation models, our GPG2A can synthesize highly realistic images.

To further improve the synthesis quality and fuse surrounding information not fully represented in the BEV layout map, such as block types (e.g., commercial or residential), we obtain ground image descriptions from large language models (e.g. Gemini). These descriptions are fed into ControlNet [57] alongside the BEV layout maps, as shown in Fig. 1. Our research not only addresses G2A synthesis but also proposes the VIGORv2 dataset, which includes center-aligned aerial-ground image pairs, layout maps, and text descriptions of ground images to train our GPG2A.

Moreover, we illustrate the practical value of GPG2A, specifically in two downstream applications, 1) data augmentation for CVGL, and 2) sketch-based region search. We show that synthesized data from our GPG2A can enhance the performance of existing CVGL models. Additionally, we illustrate the potential of synthesized images in sketch-based image retrieval, providing a more explainable and controllable approach. By presenting GPG2A, VIGORv2, and its applications, we aim to attract more researchers to advance this important and challenging field.

Our contribution can be summarized in three-folds,

- We propose GPG2A, a novel two-stage model that tackles the G2A image synthesis task. The first stage explicitly preserves the geometric layout by predicting the BEV maps from ground images. The second stage synthesizes aerial images by conditioning on the layout maps and text prompts of the ground images by using a diffusion model.
- We put forward a novel multi-modal cross-view dataset, namely, VIGORv2. Upon the existing VIGOR [64] dataset, we collected center-aligned aerial images, BEV layout maps, and text descriptions of ground images. VIGORv2 is the first cross-view dataset with image, text, and map modalities.
- We evaluate our GPG2A by using SOTA CVGL models and a customized FID [17] score. Extensive experiments demonstrate the outstanding performance of the proposed GPG2A on both same-area and cross-area protocols of VIGORv2. Moreover, the proposed approach paves the way for many applications. We demonstrate two downstream applications of our GPG2A: 1) Data augmentation for CVGL and 2) Sketch-based Region Search.

2. Related Work

Cross-View Image Synthesis: Regmi *et al*. [31] introduced cross-view image synthesis, dividing it into two sub-tasks: Aerial-to-Ground (A2G) and Ground-to-Aerial (G2A) synthesis. A2G synthesizes ground images from aerial images, while G2A tackles the inverse problem. Regmi *et al*. [31] tackled these two tasks by conditional GANs [15]. Another GAN-based approach in [44] conditions on segmentation maps of the target view, providing strong geometric prior assumptions.

Recently, the A2G task has been actively studied further by enhancing GANs [51], with CVGL [47], and leveraging geometric priors [25, 40]. Some more recent papers [12, 39] tackled satellite image synthesis from maps using diffusion models. However, G2A remains less explored or often simplified by assuming strong priors as conditional inputs. This lack of research is attributed to the inherent challenges of the G2A task, such as occlusions and the limited resolution of objects in the ground images.

The most relevant research field to this work is Bird's Eye View (BEV) prediction which aims to predict overhead segmentation from ground views. Most BEV studies are designed for autonomous driving [19, 36, 37, 43, 54, 62] that focus on closer objects such as vehicles or pedestrians. In contrast, we aim to predict the BEV map by focusing on distant objects like buildings and roads, which are typically more than 30 meters away. Therefore, due to the longer distances and the distortion in view transformation, the existing BEV methods are not preferred for this task.

Inspired by the recent success of diffusion models [27, 34, 57] in various tasks [9, 18, 26, 30, 55], we propose GPG2A which is a novel two-stage model to solve the G2A image synthesis problem. GPG2A closes the domain gap between the aerial and ground views by introducing an intermediate BEV layout stage. Our comprehensive experiments demonstrate that this innovative approach remarkably enhances the quality of the synthesized aerial image.

Cross-View Datasets: Cross-view geo-localization and cross-view synthesis share many common attributes. Therefore, these two tasks are usually conducted on the same

datasets [23, 50, 61, 64]. However, none of these datasets meet the requirement of our GPG2A, since the absence of corresponding layout maps and text description of ground images. Additionally, some datasets are unsuitable for real scenarios because they lack complex scenes [49]. For example, CVUSA [50] collects images from rural areas in the U.S. CVACT [23] only contains images from one single city in Australia. The images in University-1652 [61] are exclusively for campus buildings. Fortunately, VIGOR [64] collected aerial and ground images in four major U.S. cities. However, VIGOR is designed for the many-to-one CVGL task, resulting in the misalignment of the aerial-ground image pairs. This misalignment reduces the co-visibility between the ground and aerial views, making it unsuitable for the G2A task. To this end, we propose VIGORv2 to accommodate the needs of G2A synthesis. Our proposed dataset will be publicly available for further research.

3. VIGORv2 Dataset

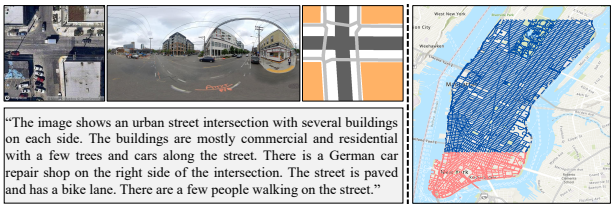

Figure 2. Left: Aerial image (left), ground image (middle), BEV layout map (right), and text description (bottom) from VIGORv2. Right: The new training (blue lines) and testing (red lines) geographically split the New York City portion of VIGORv2. The non-overlapping training and testing sets prevent data leakage.

As mentioned in Sec. 1, we propose VIGORv2 to accommodate the needs of the G2A image synthesis task. Our solution involves retaining the ground images from VIGOR while re-collecting center-aligned aerial images. In addition to the newly collected aerial images, we enhance the VIGOR dataset by introducing two new modalities: BEV layout maps and text descriptions of ground images. These additional modalities provide rich spatial contextual information and descriptive fine-grained details from the text, resulting in a more robust and comprehensive dataset. Our BEV layout maps offer much more accuracy and contain more classes than previous work [31] which uses off-the-shelf segmentation models [22].

Aerial Imagery: For each ground image, we first extract its latitude and longitude and then request an aerial image centered on this location from MapBox [1] API with a resolution of 300×300 and a zoom level of 18.5. We empirically chose this zoom level by visually inspecting that the aerial image covers most of the visual areas on ground images.

BEV Layout Maps: Accurate BEV Layout Maps are needed to train our GPG2A. Inspired by recent work [38], we collect BEV maps through OpenStreetMap [28] API with the location of the ground image and a zoom level similar to 18.5. Specifically, we select 7 most frequent categories to render with different colors in the BEV layout map: building, parking, playground, forest, water, path, and road. The rendered BEV layout map shares the same resolution as the aerial image as of 300×300.

	VIGOR [64]	VIGORv2 (ours)
Ground Images	105,214	105,214
Aerial Images	90,618	105,214
Layout Maps	N/A	105,214
Geographically Splits	×	✓
Text Description of Ground Image	×	✓
Words per Description	N/A	49.82

Table 1. Statistics comparison between the original VIGOR [64] datasets and our proposed VIGORv2.

Text Descriptions: Surrounding environment information such as the types of blocks and texture of buildings is valuable in G2A image synthesis. In our GPG2A, the text description is assigned to convey such information to the model. To this end, we utilize Google's Gemini [46] to generate the text descriptions. Gemini [46] is an easy-to-access and accurate LLM that can be utilized as an image-to-text model to describe the ground images. We used the Google Gemini API[1], with two inputs: the ground image and a custom-designed prompt. ***For more details of the prompts, please refer to the supplementary material.*** A randomly sampled image-text pair is shown in Fig. 2.

Geographical Dataset Splits: One challenge in applying the original VIGOR on the G2A task is data leakage. This leakage is caused by the overlap between the training and testing data, i.e., samples from both sets were captured nearby, often along shared streets. To tackle this issue, we adopt a train-test split based on the geographic location of an image within the city. Specifically, we divide each city into northern and southern regions. The north, covering 80% of the city, is designated for training, while the remaining 20% in the south is allocated for testing. Fig. 2 visualizes the new training and testing splits in New York City. Moreover, following the original VIGOR dataset, we also established **same-area** (training on 4 cities and testing on 4 cities) and **cross-area** (training on Seattle and New York, testing on San Francisco and Chicago) protocols for comprehensive evaluation purposes. A comparison between the original VIGOR [64] and our VIGORv2 is summarized in Tab. 1. ***For more details regarding our VIGORv2, please refer to our supplementary material.***

4. Methodology

Considering a center-aligned ground-aerial image pair I_g and I_a, GPG2A learns the transformation from the ground view to the aerial view through generating an aerial image $\hat{I}_a$ from I_g. Directly learning this transformation while preserving visual and geometrical information is challenging, primarily due to the significant change in viewing angle be-

[1] https://ai.google.dev/tutorials/python_quickstart

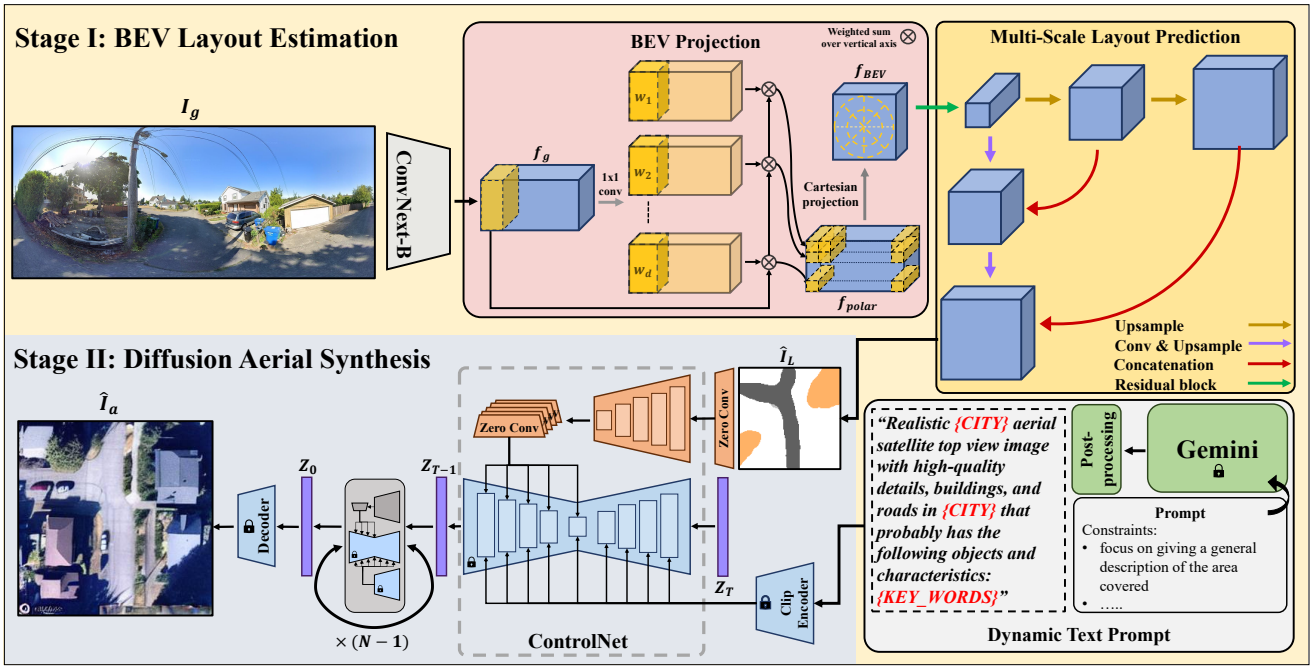

Figure 3. The main architecture of our GPG2A. The first stage is composed of BEV projection and multi-scale layout prediction. Each column in f_g is projected into a polar ray in f_{BEV}. The multi-scale network generates the BEV layout map. Then, the second stage synthesizes the aerial image using both $\hat{I}_L$ and the dynamic text prompt. All blocks with a lock symbol indicate a frozen model

tween ground and aerial perspectives. To address this challenge, we hypothesize that conditioning geometric priors as an intermediate step improves the synthesis process. Thus, we propose a two-stage model that synthesizes $\hat{I}_a$ by explicitly learning the geometry of I_a from estimating a BEV layout map $\hat{I}_L$ from I_g. Solely depending on the spatial geometric cues from $\hat{I}_L$ would miss textural details. To address this, we incorporate text descriptions of the ground image to complement $\hat{I}_L$. These text descriptions are rich in conditioning information that adds realism and fidelity to the generated aerial image.

Our GPG2A model can be formalized as follows,

$$\hat{I}_a = f\left(h_\phi(I_g),\ \tau(I_g)\right) \tag{1}$$

In Eq. (1), the first stage (BEV Layout Estimation) h is parameterized by ϕ, in which a BEV layout map is estimated from the given ground image I_g. This layout is expected to share the geometry of I_a. τ is a text extraction module that generates the text description of I_g. The second stage (Diffusion Aerial Synthesis), f, is a pre-trained ControlNet [57] model where we condition the estimated BEV layout in addition to the extracted text description from I_g.

4.1. Stage I: BEV Layout Estimation

The first stage of GPG2A estimates the BEV layout map $\hat{I}_L$ from the input ground image I_g. Initially, the ground image undergoes processing through a backbone network, which extracts a latent representation denoted as $f_g \in \mathbb{R}^{c\times h\times w}$, where c, h, and w are the channel, height, and width dimensions, respectively. For this work, we adopt ConvNeXt-B [24] as our backbone network. Subsequently, we derive a BEV feature map by projecting f_g into the polar space. This BEV feature gets decoded to produce the segmentation layout map $\hat{I}_L$.

Polar transformations have recently found success in both geo-localization [41] and BEV estimation [13, 37, 38]. Therefore, we aim to transform f_g into the polar feature representation $f_{polar} \in \mathbb{R}^{c\times d\times w}$, where d is the introduced depth dimension. f_{polar} maps each column in f_g into a polar ray of d cells. Each cell is a result of a dynamic weighted average of its corresponding column in f_g. These dynamic weights are introduced by expanding f_g along the depth dimension using 1×1 convolutions, followed by softmax normalization. Thus, as each column in f_g is dynamically weighted d times to produce d cells in the polar ray, we establish a dynamic learnable depth-aware representation of f_g, denoted as f_{polar}, as visualized in Stage I in Fig. 3.

To formalize the dynamic polar projection, we define the dynamic weights as $W_{depth} = g_\theta(I_g) \in \mathbb{R}^{c\times d\times h\times w}$, where g represents the 1×1 convolution network parameterized by θ, which expands f_g along the new depth dimension. To compute the weighted average for all columns in f_g, we compute the element-wise multiplication of f_g and W_{depth} for each d in the depth dimension, by splitting W_{depth} into d matrices of shape $[c\times h\times w]$. Subsequently, we sum over the h dimension and concatenate all d multiplication results to obtain f_{polar} with shape $[c\times d\times w]$. The extraction of f_{polar} can be formulated as follows,

$$f_{polar} = \sum_h (f_g \cdot \sigma(W_{depth})) \quad \forall d_i \in d, \tag{2}$$

where σ is softmax normalization along the h dimension,

and the operation is done for all elements d_i in the d dimension. f_{polar} is then transformed into $f_{BEV} \in \mathbb{R}^{c \times k \times k}$ where $k \in \mathbb{Z}^+$ by,

$$r_{BEV} = \sqrt{x_{polar}^2 + y_{polar}^2}\ ,\ \theta_{BEV} \quad = tan^{-1}(\frac{y}{x}), \quad (3)$$

where (x_{polar}, y_{polar}) is any point in f_{polar} and (r_{BEV}, θ_{BEV}) is its corresponding polar coordinates in f_{BEV}. Finally, we obtain the tensor of f_{BEV} by resampling it into Cartesian coordinates.

To decode f_{BEV} into a segmentation map $\hat{I}_L$, we propose the multi-scale layout prediction module (MSLP), as illustrated in Fig. 3. The decoding network is composed of a residual block [16] followed by a multi-scale feature concatenation structure. The residual block is composed of two convolution layers and a skip-connection to process and filter f_{BEV} before upsampling it into the pixel space. In MSLP, f_{BEV} is upsampled by concatenating signals from two network branches. Both branches are bilinear upsamplers but one is with additional convolution layers. This design simultaneously refines and upsamples the processed BEV feature map by learning both low- and high-level semantic information. To train stage I, we adopt the Dice loss [42] defined as, $L_{Dice} = 1 - \frac{2\,|\hat{I}_L \cap I_L|}{|\hat{I}_L| + |I_L|}$.

4.2. Stage II: Diffusion Aerial Synthesis

4.2.1 ControlNet

ControlNet [57] adds spatial conditioning to pre-trained text-to-image diffusion models [34] by utilizing zero-convolution layers. Its promising results show that it can generate realistic images in multiple domains [14]. In stage II, we condition on both the predicted layout map $\hat{I}_L$ and text prompts from our text extraction module τ. The first carries the spatial and geometric priors, while the latter introduces textural consistency with the aerial image I_a.

4.2.2 Dynamic Text Prompts

We leverage the versatility of the diffusion model by incorporating an additional modality, specifically text conditions. These encapsulate the environmental and scenic context of the captured area, improving the synthesized aerial images with elements beyond geometry. However, the raw Gemini descriptions contain minor errors and hallucinations which eventually degrade the quality of the generated aerial images (see Sec. 5.4).

We employ a text extraction post-processing that filters the text and extracts keywords of interest. The extracted keywords, as well as the prior knowledge, e.g., city name, are combined in a template (see Fig. 3 "Dynamic Text Prompt" panel for details). We focus only on important details in the raw description by constraining it in the template, naming this process, the "dynamic" text prompt. To perform keyword extraction, we adopt a BERT-based off-the-shelf model[2] which utilizes BERT embedding and cosine similarity to identify m N-gram phrases that closely resemble the raw text. The key phrases are ranked by the Maximal Marginal Relevance (MMR) technique [7] based on their relevance to the text. ***Refer to our supplementary material for more information about MMR.***

4.2.3 Model Training

In GPG2A's second stage, the following simplified objective from [27] is used to train the diffusion model.

$$L_{LDM} \coloneqq \mathbb{E}_{\mathbf{Z}, \epsilon \sim \mathcal{N}(0,1), t}\Big[\|\epsilon - \epsilon_\theta(t, z_t, \tau(I_g), \hat{I}_L)\|_2^2\Big], \quad (4)$$

where $\mathbf{Z}$ is the latent representation of the images generated from the pre-trained Variational Autoencoder (VAE) from [34]. ϵ_θ is parameterized by θ and defined as the time-conditioned U-net [35] with our additions, i.e., text extraction module τ and the BEV layout map $\hat{I}_L$. t is the time step value in the diffusion process.

5. Experiments

5.1. Evaluation metrics

Figure 4. One reference image and three samples for evaluation metrics comparisons.

Metrics	Sample 1	Sample 2	Sample 3
PSNR ↑	8.587	8.302	8.554
SSIM ↓	0.062	0.052	0.060
LPIPS ↓	0.753	0.722	0.784
Sim_s ↓	0.510	0.362	0.409
Sim_c ↓	0.467	0.370	0.416

Table 2. Evaluation metrics comparison between the sample images and the reference in Fig. 4. Existing methods (PSNR, SSIM, LPIPS) can hardly capture the similarity in aerial images. ↑ means higher better. ↓ means lower better.

In the G2A task, popular image quality evaluation metrics such as PSNR, SSIM [48], and LPIPS [58] are insufficient to evaluate the similarity in aerial images. For illustration, we select one reference image and three test images as shown in Fig. 4. Sample 1 shares a different layout (horizontal street) than samples 2 and 3 (vertical street). We measure the similarity between each sample and the reference image using the aforementioned metrics in Tab. 2. PSNR, SSIM, and LPIPS do not reflect the similarities as all three

[2] https://maartengr.github.io/KeyBERT/

metrics show minor differences. This is because these metrics either only estimate pixel-level similarity (PSNR and SSIM) or lack knowledge of aerial image data (LPIPS).

To address the above-mentioned issue, we propose a new approach to evaluate our proposed methods by using one of the state-of-the-art cross-view geo-localization (CVGL) model [41] to estimate the similarity between real and synthesized aerial images. The goal of CVGL is to minimize the distance between matched aerial-ground pairs and maximize the distance between the unmatched ones. Formally, denote f^a, f^g, and $\hat{f}^a$ as the L_2 normalized features for real aerial images, corresponding ground images, and synthesized aerial images, respectively from a well-trained CVGL model (i.e. SAFA [41]). If the synthesized aerial image is realistic and geometrically preserved, the distance between f^a and $\hat{f}^a$ should be small and we name it same-view similarity metric (Sim_s) which are formally defined as follows,

$$Sim_s = \frac{1}{N}\sum_{i=1}^{N}\frac{2 - 2\times(f^a \cdot \hat{f}^a)}{4}, \tag{5}$$

where N is the number of samples. Correspondingly, we also evaluate the similarity between f^g and $\hat{f}^a$ and we name it cross-view similarity metric (Sim_c) which can be easily obtained by replacing f^a into f^g in Eq. (5). To extract the features, we train the SAFA [41] on the training set of VIGORv2. In Tab. 2, Sim_s and Sim_c shows that sample 2 and sample 3 are closer to the reference image than sample 1. This indicates its efficacy in evaluating the synthesized images in this task. Besides Sim_s and Sim_c, we also adopt a customized FID [17] score, namely FID_{SAFA} that leverages the features (f^a and $\hat{f}^a$) to evaluate the divergence between real images and synthesized images. ***For more details, please refer to the supplementary material.***

5.2. Quantitative Results

Method	Same-area			Cross-area		
	$Sim_s \downarrow$	$Sim_c \downarrow$	$\text{FID}_{\text{SAFA}} \downarrow$	$Sim_s \downarrow$	$Sim_c \downarrow$	$\text{FID}_{\text{SAFA}} \downarrow$
X-seq	0.392	0.438	0.411	0.392	0.454	0.570
X-fork	0.341	0.423	0.151	0.372	0.445	0.357
ControlNet†	0.435	0.415	0.154	0.446	0.405	0.386
ControlNet‡	0.369	0.412	0.110	0.409	0.420	0.220
GPG2A (ours)	0.295	0.402	0.079	0.333	0.392	0.197

Table 3. Same-area and cross-area benchmark results between our proposed GPG2A with baseline methods on our VIGORv2. The best results are highlighted in a gray background. † indicates that a fixed text prompt is used for training the ControlNet. ‡ indicates training the ControlNet with the dynamic text prompts proposed in this paper. ↓ indicates that the lower value is better.

To evaluate our GPG2A, we benchmark it on the proposed VIGORv2 dataset in both same-area and cross-area protocols. As discussed in Sec. 5.1, we rely on the Sim_c, Sim_s, and FID score for comparison. We choose ControlNet [57], X-fork, and X-seq [31] as the baseline methods. For a fair comparison, two versions of ControlNet were evaluated, one with a constant prompt condition and another with our proposed dynamic prompt. To our best knowledge, X-fork and X-seq are the only models to tackle the G2A task. The experimental results are presented in Tab. 3 in which the left panel shows the same-area results and the right panel shows the cross-area results. Our proposed GPG2A achieves the best results among all the baseline methods in both same-area and cross-area experiments. Notably, the Sim_s and FID_{SAFA} of our GPG2A are substantially better than other baseline methods. On the other hand, ControlNet [57] does not outperform the GAN-based X-fork [31] in Sim_s and Sim_c. This illustrates that without the input of the geometric prior, i.e., BEV layout maps, ControlNet can hardly infer the ground to aerial view changes. This observation supports our two-stage pipeline which divides the BEV estimation and aerial synthesis. A clear improvement in ControlNet can be noticed when using the dynamic text prompt, which validates the use of our text extraction module. In the cross-area experiment, we notice that X-seq has a larger Sim_c and FID_{SAFA} score. This might be attributed to the overfitting issue in this GAN-based method that cannot generalize to unseen data. However, our proposed GPG2A can still maintain an outstanding performance in the cross-area experiments. ***For conventional PSNR, SSIM, and LPIPS scores, please refer to our supplementary materials.***

5.3. Qualitative Results

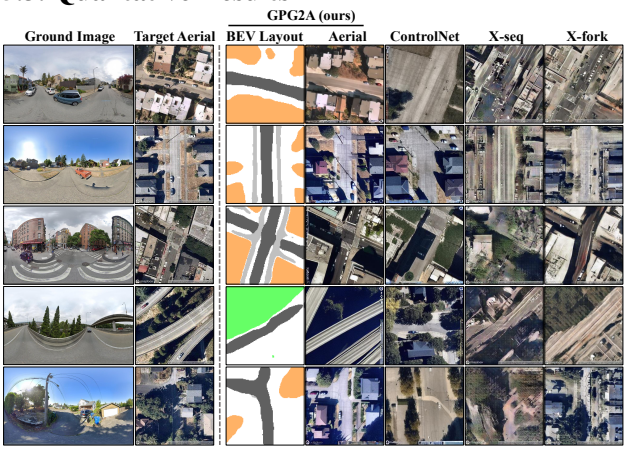

Figure 5. Same-area qualitative comparison. From left to right are ground images, target aerial images, ours synthesized BEV layouts and aerial images, ControlNet [57], X-seq [31], and X-fork [31].

Same-Area Experiment: Some randomly selected samples are visualized in Fig. 5. For our GPG2A, we present both generated aerial images and predicted BEV layout maps. Notably, the synthesized aerial images and BEV layout maps share geometric structures, providing empirical support to our hypothesis that the BEV map prior would lead to better synthesis. Compared to other baselines, especially in the first and the fourth example in Fig. 5, GPG2A

preserves geometry and generates high-quality aerial images with details. However, ControlNet has some details, e.g., roads and trees, but lacks geometric correspondence. X-fork and X-seq generate blurry images without details.

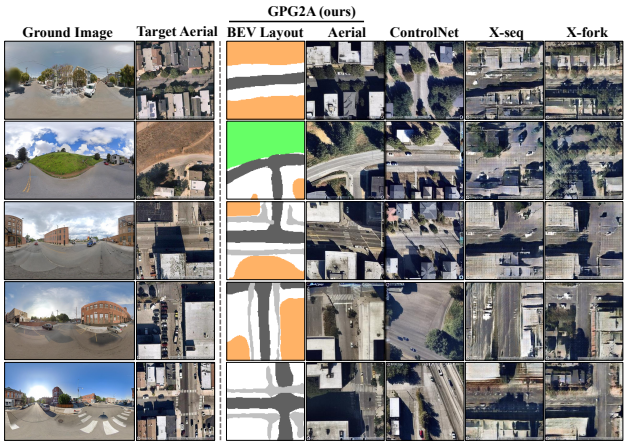

Figure 6. Cross-area qualitative comparison. From left to right are ground images, target aerial images, ours synthesized BEV layouts and aerial images, ControlNet [57], X-seq [31], and X-fork [31].

Cross-Area Experiment: To further validate the generalization of GPG2A on unseen data, we devise a cross-area experiment as visualized in Fig. 6. It is clear to see that the accurate estimation of the BEV layout maps preserves geometric consistency even in unseen scenarios. To be noticed, some disparities appear in environmental details, such as the appearance of buildings (the fifth example in Fig. 6). On the other hand, all other baseline methods generated samples lacking both geometry and details compared with our GPG2A. ***For more visualizations with different conditions and failure cases, please refer to supplementary materials.***

5.4. Ablation Studies

Prompt	Same-area			Cross-area		
	$Sim_s \downarrow$	$Sim_c \downarrow$	$\text{FID}_{\text{SAFA}} \downarrow$	$Sim_s \downarrow$	$Sim_c \downarrow$	$\text{FID}_{\text{SAFA}} \downarrow$
Raw	0.383	0.425	0.123	0.384	0.412	0.227
Constant	0.323	0.418	0.131	0.362	0.407	0.259
City-only	0.316	0.419	0.087	0.356	0.424	0.208
Dynamic	0.295	0.402	0.079	0.333	0.392	0.197

Table 4. Ablation study of the text prompt in the proposed GPG2A. 'Constant' indicates fixing the text prompt. 'Raw' stands for using raw text descriptions from Gemini without keyword selection. 'City-only' means varying the city name in the prompt. 'Dynamic' stands for the proposed dynamic text prompt.

Text prompt: Text prompts provide important contextual details for GPG2A, as mentioned in Sec. 4.2. In this experiment, we ablate different types of prompts in training to demonstrate the effectiveness of our dynamic prompt. We study three additional prompts: the constant prompt, a fixed generic text prompt; the raw prompt, which directly applies the Gemini output; and the city-only prompt, which only varies the city name. The experiment results are presented in Tab. 4. First, the "Raw" prompt has the worst results due to the lengthy text from Gemini (potentially with hallucination), resulting in a noisy signal to the model. It is noted that the "constant" prompt (similar to an empty prompt because both embedding values never change during training) is better than the "Raw" prompt in Sim_s and Sim_c but worse in FID_{SAFA}. This degradation might be attributed to the absence of ground surrounding information. This also reveals the importance of our dynamic prompt which boosts the model in both same-area and cross-area settings.

Method	$Sim_s \downarrow$	$Sim_c \downarrow$	$\text{FID}_{\text{SAFA}} \downarrow$
w/o MSLP	0.465	0.478	0.426
w/o Stage I	0.435	0.415	0.154
GPG2A (ours)	0.295	0.402	0.079

Table 5. Ablation study on the effectiveness of our GPG2A stage I. 'MSLP' stands for the multi-scale layout prediction module. To remove MSLP, we input f_{bev} to stage II. To remove stage I, we input the ground image directly to stage II.

Effectiveness of Stage I: To verify our two-stage design in GPG2A, we conduct two ablation studies, 1) removing the multi-scale layout prediction module in stage I by using f_{bev} in stage II directly, and 2) removing the stage I completely by conditioning directly on ground images. As indicated in Tab. 5, our proposed GPG2A is constantly better than these two variants which is a firm support to our assumption that the intermediate BEV layout alleviates the challenge in G2A transformation. We also notice that conditioning ground image directly is better than using f_{bev}. As f_{bev} only contains latent polar features which can hardly be utilized in stage II synthesis. ***For more ablation studies, please refer to the supplementary material.***

6. Applications

To further evaluate our GPG2A model and validate G2A image synthesis across domains, we apply it to two real-world applications: 1) Data augmentation for geo-localization and 2) Sketch-based region search.

6.1. Data Augmentation for Geo-localization

	$p_\circ$	R@1 ↑	R@5↑	R@10↑	R@1%↑
same-area	0	83.58%	93.72%	96.08%	99.27%
	0.4	86.33%	94.76%	96.27%	**99.38%**
	0.6	**86.84%**	**94.98%**	96.30%	99.34%
	0.8	85.77%	94.71%	**96.44%**	99.30%
cross-area	0	50.03%	70.17%	77.70%	94.03%
	0.4	51.55%	71.96%	78.18%	94.10%
	0.6	**52.84%**	**72.32%**	**78.45%**	**94.19%**
	0.8	50.11%	70.98%	77.60%	93.99%

Table 6. Results of our data augmentation on SAFA. $p_\circ = 0$ indicates no augmentation is applied. It is noticeable that our augmentation can improve in both same-area and cross-area performance.

Many augmentation techniques have been proposed [8, 33, 49, 59] to train robust CVGL models. In this application, we propose to apply aerial images generated by GPG2A in the Mixup augmentation method [56] to train robust CVGL models. The mixup augmentation can be defined as follows,

$$\hat{x} = \begin{cases} \lambda\, x_{fake} + (1-\lambda)\, x_{real} & p \leq p_\circ \\ x_{real} & p > p_\circ, \end{cases} \tag{6}$$

where $\hat{x}$, x_{fake}, and x_{real} are the augmented, generated (GPG2A output), and real images, respectively. λ is the mixup strength. $p_\circ$ is the probability of applying the augmentation. We apply this augmentation to the well-known SAFA [41] model which was trained in New York and Seattle, and evaluated in both same-area and cross-area settings.

Tab. 6 shows the performance of our data augmentation in the same-area and cross-area settings. We evaluate the performance using recall accuracy at top K (R@K) [41], which measures the likelihood that the ground truth aerial image ranks within the top K predictions. Overall, the proposed augmentation improved performance across all metrics. Specifically, it brings 3.26% and 2.81% improvements to same-area and cross-area tests respectively while $p_\circ = 0.6$ on R@1. We also notice that the performance decreases while $p = 0.8$. This might indicate the lack of convergence of the model because of the stronger augmentation. ***Please refer to our supplementary material for experiments on more recent CVGL models.***

6.2. Sketch-based Region Search

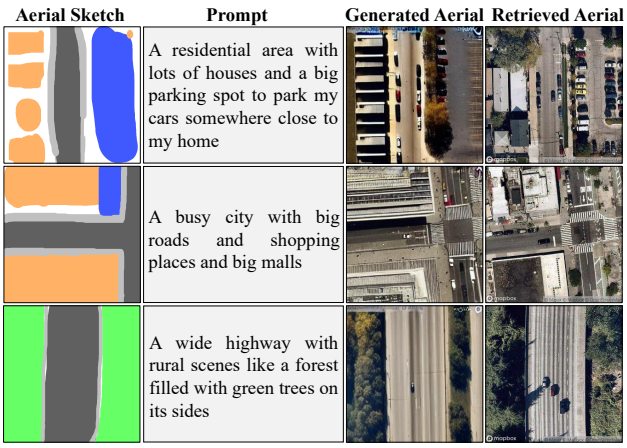

Figure 7. Synthesis and retrieval results of the sketch-based region search application. Each color in the layout sketch represents a class as follows: orange, black, grey, blue, and green reflect buildings, streets, sidewalks, parking lots, and trees, respectively.

Aerial image search is one of the most challenging tasks in remote sensing [63], particularly when a query image is represented by mental maps, such as hand-drawing sketches [53] without low-level details. In this task, we assume that a hand-drawn sketch and a text description of the surrounding environment are given. The goal is to find similar aerial images from an imagery database. In this way, the user can find points of interest by using only sketches and descriptions. We use the second stage of GPG2A to synthesize a fake aerial image from the sketch and the text description. Then, we retrieve the most similar aerial image from the reference database by calculating the closest latent features in Euclidean distance. To achieve this, a pre-trained SAFA model is adopted to extract latent features.

Fig. 7 illustrates 3 retrieval results with different scenery and objects. For example, a parking lot was in the first sample, while a highway was included in the third sample. Both synthesized and retrieved images showed strong correspondence with the given sketches and descriptions. To further evaluate this pipeline, we conducted a survey that asked 61 volunteers to identify similarities between 5 groups of the input (aerial sketch and text prompt) with three different aerial images (corresponding top-1 retrieved aerial image, the 5th retrieved aerial image, and a random aerial image). The results show that 66% of the volunteers believe the top-1 retrieved images correspond to the input aerial sketch and text prompt. This number drops to 60% in the 5th retrieved aerial image. While only 24% of the people think random aerial images are similar to the input. It is noteworthy that visualization of generated aerial images from sketches boosts search explainability. Most previous works [53, 63] aim to find a common latent space between sketches and aerial images, which lacks interpretability. ***For more details, please refer to supplementary material.***

7. Conclusion and Future Works

In this paper, we propose GPG2A which is a two-stage model that generates geometry-preserved aerial images from ground images by conditioning on predicted BEV layouts and text descriptions. To alleviate the problem of lacking datasets for benchmarking, we propose VIGORv2, which is built upon the VIGOR [64] dataset with newly collected aerial images, BEV layout maps, and text descriptions. Our GPG2A outperforms existing baselines on the VIGORv2 dataset. Additionally, we apply our GPG2A on two downstream tasks to show its potential application.

As a novel research field, there are many opportunities to advance this research such as feature fusion from ground videos and fine-grained conditioning techniques to generate more realistic and diverse aerial images.

8. Acknowledgement

This work was supported by the National Science Foundation under Grants No. 2218063. Computations were performed at the Vermont Advanced Computing Center (VACC), which was supported by AMD's donation of critical hardware and resources from its HPC Fund.

References

[1] Mapbox. https://www.mapbox.com/. 3

[2] Mapillary. https://www.mapillary.com/app/. 1

[3] New york state gis resources. https://gis.ny.gov/orthoimagery. 1

[4] Abolfazl Abdollahi and Biswajeet Pradhan. Urban vegetation mapping from aerial imagery using explainable ai (xai). *Sensors*, 21(14):4738, 2021. 1

[5] Jacob Levy Abitbol and Marton Karsai. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence*, 2(11):684–692, 2020. 1

[6] Josef Aschbacher and Maria Pilar Milagro-Pérez. The european earth monitoring (gmes) programme: Status and perspectives. *Remote Sensing of Environment*, 120:3–8, 2012. 1

[7] Kamil Bennani-Smires, Claudiu Musat, Andreea Hossmann, Michael Baeriswyl, and Martin Jaggi. Simple unsupervised keyphrase extraction using sentence embeddings. *arXiv preprint arXiv:1801.04470*, 2018. 5

[8] Gabriele Berton, Riccardo Mereu, Gabriele Trivigno, Carlo Masone, Gabriela Csurka, Torsten Sattler, and Barbara Caputo. Deep visual geo-localization benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5396–5407, June 2022. 8

[9] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5968–5976, 2023. 2

[10] Benjamin Coifman, Mark McCord, Rabi G Mishalani, and Keith Redmill. Surface transportation surveillance from unmanned aerial vehicles. In *Proc. of the 83rd Annual Meeting of the Transportation Research Board*, volume 28, 2004. 1

[11] Julien Cornebise, Ivan Oršolić, and Freddie Kalaitzis. Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution. *Advances in Neural Information Processing Systems*, 35:25979–25991, 2022. 1

[12] Miguel Espinosa and Elliot J Crowley. Generate your own scotland: Satellite image generation conditioned on maps. *arXiv preprint arXiv:2308.16648*, 2023. 2

[13] Florian Fervers, Sebastian Bullinger, Christoph Bodensteiner, Michael Arens, and Rainer Stiefelhagen. C-bev: Contrastive bird's eye view training for cross-view image retrieval and 3-dof pose estimation, 2023. 4

[14] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601*, 2023. 5

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[17] 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. *Advances in neural information processing systems*, 30, 2017. 2, 6

[18] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[19] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15273–15282, October 2021. 2

[20] James R Irons, John L Dwyer, and Julia A Barsi. The next landsat satellite: The landsat data continuity mission. *Remote sensing of environment*, 122:11–21, 2012. 1

[21] Ruimin Ke, Zhibin Li, Jinjun Tang, Zewen Pan, and Yinhai Wang. Real-time traffic flow parameter estimation from uav video based on ensemble classifier and optical flow. *IEEE Transactions on Intelligent Transportation Systems*, 20(1):54–64, 2018. 1

[22] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1925–1934, 2017. 3

[23] 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 (CVPR)*, June 2019. 3

[24] 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 (CVPR)*, pages 11976–11986, June 2022. 4

[25] 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. 2

[26] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint arXiv:2302.08453*, 2023. 2

[27] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. 2, 5

[28] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017. 3

[29] Anuj Puri. A survey of unmanned aerial vehicles (uav) for traffic surveillance. *Department of computer science and engineering, University of South Florida*, pages 1–29, 2005. 1

[30] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022. 2

[31] 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, 6, 7

[32] M. Fasi Ur Rehman, Izza Aftab, Waqas Sultani, and Mohsen Ali. Mapping temporary slums from satellite imagery using a semi-supervised approach. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2022. 1

[33] Royston Rodrigues and Masahiro Tani. Are these from the same place? seeing the unseen in cross-view image geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 3753–3761, January 2021. 8

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 2, 5

[35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015. 5

[36] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Enabling spatio-temporal aggregation in birds-eye-view vehicle estimation. In *2021 ieee international conference on robotics and automation (icra)*, pages 5133–5139. IEEE, 2021. 2

[37] Avishkar Saha, Oscar Mendez, Chris Russell, and Richard Bowden. Translating images into maps. In *2022 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2022. 2, 4

[38] Paul-Edouard Sarlin, Daniel DeTone, Tsun-Yi Yang, Armen Avetisyan, Julian Straub, Tomasz Malisiewicz, Samuel Rota Bulò, Richard Newcombe, Peter Kontschieder, and Vasileios Balntas. Orienternet: Visual localization in 2d public maps with neural matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21632–21642, 2023. 3, 4

[39] Srikumar Sastry, Subash Khanal, Aayush Dhakal, and Nathan Jacobs. Geosynth: Contextually-aware high-resolution satellite image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 460–470, 2024. 2

[40] 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

[41] 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, 2019. 1, 4, 6, 8

[42] Carole H. Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M. Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, pages 240–248. Springer International Publishing, 2017. 5

[43] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J. Black. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13243–13252, June 2022. 2

[44] 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 (CVPR)*, June 2019. 2

[45] John R Taylor and Sarah Taylor Lovell. Mapping public and private spaces of urban agriculture in chicago through the analysis of high-resolution aerial images in google earth. *Landscape and urban planning*, 108(1):57–70, 2012. 1

[46] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 3

[47] 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

[48] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5

[49] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Image and Object Geo-Localization. *International Journal of Computer Vision*, 132(4):1350–1392, Apr. 2024. 3, 8

[50] 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 (ICCV)*, December 2015. 1, 3

[51] 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

[52] Shuo-sheng Wu, Bing Xu, and Le Wang. Urban land-use classification using variogram-based analysis with an aerial photograph. *Photogrammetric Engineering & Remote Sensing*, 72(7):813–822, 2006. 1

[53] Fang Xu, Wen Yang, Tianbi Jiang, Shijie Lin, Hao Luo, and Gui-Song Xia. Mental retrieval of remote sensing images

via adversarial sketch-image feature learning. *IEEE Transactions on Geoscience and Remote Sensing*, 58(11):7801–7814, 2020. 8

[54] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17830–17839, June 2023. 2

[55] Dongchao Yang, Jianwei Yu, Helin Wang, Wen Wang, Chao Weng, Yuexian Zou, and Dong Yu. Diffsound: Discrete diffusion model for text-to-sound generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023. 2

[56] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 8

[57] 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*, pages 3836–3847, 2023. 2, 4, 5, 6, 7

[58] 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. 5

[59] Xiaohan Zhang, Xingyu Li, Waqas Sultani, Chen Chen, and Safwan Wshah. Geodtr+: Toward generic cross-view geolocalization via geometric disentanglement. *arXiv preprint arXiv:2308.09624*, 2023. 1, 8

[60] 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*, volume 37, pages 3480–3488, 2023. 1

[61] Zhedong Zheng, Yunchao Wei, and Yi Yang. University-1652: A multi-view multi-source benchmark for drone-based geo-localization. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1395–1403, New York, NY, USA, 2020. Association for Computing Machinery. 3

[62] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *CVPR*, 2022. 2

[63] Weixun Zhou, Shawn Newsam, Congmin Li, and Zhenfeng Shao. Patternnet: A benchmark dataset for performance evaluation of remote sensing image retrieval. *ISPRS Journal of Photogrammetry and Remote Sensing*, 145:197–209, 2018. Deep Learning RS Data. 8

[64] 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 (CVPR)*, pages 3640–3649, June 2021. 1, 2, 3, 8