

Supplementary Material

Ahmad Arrabi^{1†}, Xiaohan Zhang^{1†}, Waqas Sultani², Chen Chen³, Safwan Wshah^{1*}

¹ Vermont Artificial Intelligence Lab, Department of Computer Science, University of Vermont

² Intelligent Machines Lab, Information Technology University

³Center for Research in Computer Vision, University of Central Florida

[†] These authors contributed equally. ^{*} Corresponding and senior author.

1. Introduction

In this supplementary material, we provide more information on our implementation details (Sec. 2). As mentioned in our main script, we presented the evaluation results on traditional SSIM and PSNR evaluation metrics (Sec. 3), and gave more details on the FID_{SAFA} score (Sec. 4). Furthermore, we perform an extra ablation study on limited FOV ground images (Sec. 5). We also present more qualitative samples of our proposed GPG2A in the same-area and cross-area experiments (Sec. 6). Additionally, we analyze some failure synthesis cases (Sec. 7). More aerial, ground, text, and BEV layout map samples from our VIGORv2 dataset are presented (Sec. 8) along with the description of the text data used in this research (Sec. 9) with an ablation study of visualizing the generated aerial image when varying the input text. Additionally, we show the significance of certain classes of the BEV layout map, and how they affect the generated images (Sec. 10). We also provide additional experiments, results, and more details of the two downstream applications (Sec. 11 and Sec. 12). Finally, we discuss the limitations (Sec. 13) and the societal impact (Sec. 14) of our paper.

2. Implementation Details

We implemented the GPG2A in Pytorch [11]. In stage I, we trained the model with a batch size of 128 and Adam [9] optimizer with the learning rate set to 0.0001. In the BEV projection step, the depth dimension d was set to 64, and in the Cartesian projection, k was set to 32. Thus, the polar feature f_{polar} was resampled into a Cartesian grid to derive $f_{BEV} \in \mathbb{R}^{c \times k \times k}$ where c is the latent channel dimension and we set it to 256 in our experiments. We use the Torchvision¹ official implementation of ConvNext-B [10] with ImageNet [6] pre-trained weights for the backbone feature extractor. The output of stage I is rendered in pixel space as input for stage II. In stage II, we use the Hugging Face Dif-

fuser library’s implementation of ControlNet from². To train the model, we used Adam [9] optimizer with a learning rate of 0.0001 and a batch size of 4. Stage I is trained on an AMD MI50 GPU and stage II is trained on a Nvidia V100 GPU with 20 epochs and 30 epochs, respectively.

3. SSIM, PSNR, and LPIPS Quantitative Results

As mentioned in the main script, existing evaluation metrics such as PSNR, SSIM [17], and LPIPS [20] are insufficient to estimate the quality of the synthesized aerial images. Thus, we did not show the results of SSIM, PSNR, and LPIPS scores. In this section, we provide a comprehensive study that compares PSNR, SSIM, and LPIPS. The results are presented in Tab. 1. As indicated in the table, SSIM, PSNR, and LPIPS show minimal variants in both same-area and cross-area experiments, which reflects our point in Sec. 5.1 of the main script that existing evaluation metrics are insufficient to evaluate the quality of the synthesized aerial images.

4. More details about FID_{SAFA}

Similar to the original FID score [7], FID_{SAFA} leverages the features (f^a and \hat{f}^a) to evaluate the divergence between real and synthesized images. However, to better evaluate the quality of aerial images, we choose to use the features extracted by pre-trained SAFA [15] which yields better feature quality than the original one, especially for aerial images. FID_{SAFA} can be formally written as follows,

$$\text{FID}_{\text{SAFA}} = \|\mu^a - \hat{\mu}^a\| + \text{Tr}(\Sigma^a + \hat{\Sigma}^a - 2(\Sigma^a \hat{\Sigma}^a)^{\frac{1}{2}}) \quad (1)$$

where $\mathcal{N}(\mu^a, \Sigma^a)$ and $\mathcal{N}(\hat{\mu}^a, \hat{\Sigma}^a)$ are multivariate normal distributions estimate from f^a and \hat{f}^a .

¹<https://pytorch.org/vision/main/models/convnext.html>

²<https://huggingface.co/docs/diffusers/using-diffusers/controlnet>

Method	Same-area			Cross-area		
	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow	SSIM \uparrow	PSNR \uparrow	LPIPS \downarrow
X-fork [13]	0.11	10.66	0.58	0.10	10.34	0.60
X-seq [13]	0.10	10.30	0.61	0.09	10.15	0.62
ControlNet [19]	0.09	9.62	0.66	0.09	9.90	0.66
GPG2A [†]	0.12	10.40	0.63	0.09	9.65	0.60
GPG2A [‡]	0.10	9.59	0.62	0.08	9.42	0.60
GPG2A (ours)	0.10	10.07	0.62	0.09	10.17	0.61

Table 1. PSNR, SSIM [17], and LPIPS [20] results from our proposed GPG2A as well as other baselines on our proposed VIGORv2 datasets in same-area and cross-area experiments protocols. \uparrow stands for the higher value better, \downarrow stands for the lower value better. \dagger stands for constant text prompt and \ddagger stands for city text prompt.

FOV	BEV Accuracy		Synthesis Quality		
	Avg F1	mIoU	Sim_s \downarrow	Sim_c \downarrow	FID _{SAFA} \downarrow
90°	0.259	0.149	0.413	0.414	0.290
180°	0.411	0.258	0.385	0.406	0.181
270°	0.458	0.297	0.369	0.404	0.143
360°	0.565	0.394	0.295	0.402	0.079

Table 2. Ablation study on limited FOV input ground images. “BEV Accuracy” represents BEV layout prediction accuracy from Stage I and “Synthesis Quality” represents the quality of the synthesized aerial image from Stage II.

5. Limited FOV Ablation Study

In previous experiments, we assume the ground images are panoramic. In fact, limited field-of-view (FOV) images are more accessible [22]. Thus, we extend stage I of GPG2A to predict BEV layout maps from limited FOV images. As shown in Tab. 2, we use the Average F1 and IoU to evaluate the prediction accuracy in stage I, and Sim_s , Sim_c , and FID_{SAFA} to evaluate the synthesis quality in stage II. As shown in Tab. 2, with the increase of FOV, both stages improved. It is noteworthy that by comparing Tab. 2 and Table 3 of the main paper, GPG2A is still better than X-seq and X-fork with 180° FOV input in Sim_c and comparable results on Sim_s and FID_{SAFA}.

We further extend our analysis of GPG2A in limited FOV ground images by visualizing the predicted BEV layout maps in stage I. As shown in Fig. 1, the segmentation results improve as the FOV increases. The samples from the 360° FOV entail more geometric detail of the ground truth. For example, the third sample generated the full intersection while other FOVs did not. Notice how the last sample has a green segmentation class indicating trees, which is not aligned with the ground truth. However, after further investigation, it turned out that there were trees visible in both the ground and aerial counterparts as shown in Fig. 2. This indicates that stage I is dynamic and can extrapolate information from the ground image to the BEV layout.

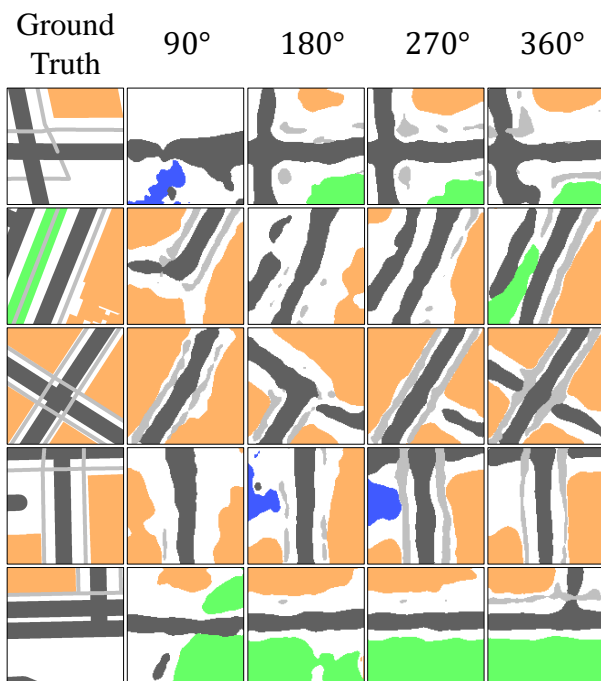


Figure 1. Visualization of the output from stage I of our GPG2A under different limited FOV input ground images. From left to right are ground truth BEV layout map, 90° FOV, 180° FOV, 270° FOV, and 360° FOV

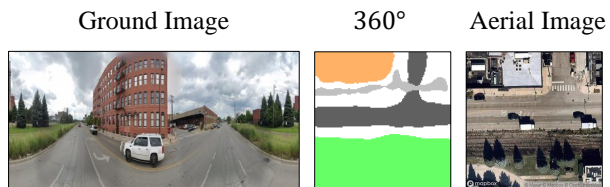


Figure 2. Test sample where stage I inferred trees from the ground image

6. More Qualitative Results

This supplementary material visualizes more diverse synthesis cases of our GPG2A in both the same- and cross-area test cases, as shown in Figs. 3 and 4, respectively. In the same-area results, we can notice how GPG2A generalizes in multiple environments, e.g., residential and urban areas. Also, GPG2A showed great attention to detail, as shown in the tree placement in the third sample, and parking synthesis in the last. Notice how the dynamic text explicitly mentioned these details which guide the synthesis. In the cross-area samples, the geometry was clearly preserved due to the accurate BEV estimates. The dynamic text carried environment details but as the model was trained on different cities, the generated images had some discrepancies with the ground truth. For example, in the first two samples, even though trees were inferred their styles did not match.

7. Failure cases

To further study the behavior of our GPG2A, we present some failure cases as shown in Fig. 5. The first example (top) keeps the geometric layout as shown in both the predicted BEV layout map and the generated aerial image. However, the generated image lacks details of the buildings and trees. This might be due to the occluded parts in the ground image as a tree was shown to cover the houses. The second (middle) and the third (bottom) examples fail to generate accurate aerial images because of some unseen objects. Specifically, there is a bridge (in the east direction of the ground image) in the ground image which is not captured by stage I of GPG2A as evidenced in the predicted layout image. Except for this bridge, the geometric layout is still preserved. For the last example, there is a two-story parking garage not visible from the ground image which causes the model to fail to capture the geometric layout. Thus, as shown in the generated aerial image, there is no corresponding geometry.

8. More Dataset Figures

In our VIGORv2, we propose a new split that maximizes the discrepancy between training and testing samples by sampling northern and southern regions within the city. We further visualize this split on all four cities in Fig. 6. Also, We provide more randomly sampled aerial, ground, text, and BEV layout samples from VIGORv2 in Fig. 7. These samples demonstrate that our newly proposed dataset, VIGORv2, contains diverse street layouts in different environments.

9. Text Data Discussion

In this section, we will give more detail about the text data used in our GPG2A. The discussion will focus on two

main points: the prompt used to collect the Gemini descriptions, and the four types of text conditions used in GPG2A (sec5.4 in the main script)

To collect the ground view descriptions, we passed the following prompt along with the ground image to Gemini:

You are an AI visual assistant who greatly understands geospatial data. Please generate a paragraph to give a high-level description of the image below with the following constraints. Your output should only be this paragraph without any introductory sentences. Please follow the following rules:

- 1. focus on giving a general description of the place in the image, don't include small-level details like pedestrians and cars.*
- 2. focus on the buildings, if any, their type, and the number of close-by buildings.*
- 3. Do not care about any weather conditions, we want a description of the geospatial area*
- 4. do not include the following words and or any similar words: 'panorama', '360', 'sky', 'car', 'truck', 'pedestrian'*
- 5. describe the environment type, for example residential, highway, urban, rural .. etc*
- 6. be limited to about 50 words maximum*
- 7. if there are people do not pay attention to them and consider them blurred out*
- 8. please only generate the output directly, do not add any introductory sentences, and only output the description paragraph*
- 9. If you have any limitations do not mention them, never talk about them*
- 10. Only output in English*

The constraints were technical to control Gemini's output and minimize hallucinations, and logical, to give accurate descriptions of the ground image, e.g., points 1, 2, 3, and 5. In point 6, we limit the number of words in the output to have consistent samples. Gemini is designed for chatbot applications, its output is often conversational with redundant greetings and introductory words. We consider these outputs as noise and ask not to include them. Lastly, we make sure to only output in English as we noticed that sometimes, for some reason, Gemini outputs Japanese words.

After collecting the ground image text descriptions utilizing the prompt above, we thoroughly analyzed our GPG2A on four different cases of text conditions. The constant text prompt is a generic prompt that describes any

Ground Image	Target Aerial Image	GPG2A (ours)		
		BEV Layout	Aerial Image	Dynamic Text
				Realistic Chicago aerial satellite top view image with high quality details with buildings and roads in Chicago that probably has the following objects and characteristics: alley surrounded residential, garbage cans alley, gray two-story house, black fence yard, white garage black
				Realistic NewYork aerial satellite top view image with high quality details with buildings and roads in NewYork that probably has the following objects and characteristics: busy urban street, street including restaurant, bakery pharmacy buildings, 57 stories tall, buildings brick 57
				Realistic NewYork aerial satellite top view image with high quality details with buildings and roads in NewYork that probably has the following objects and characteristics: entire length road, closer road buildings, road treelined median, road paved lanes, sidewalk road streetlights
				Realistic Seattle aerial satellite top view image with high quality details with buildings and roads in Seattle that probably has the following objects and characteristics: residential area houses, houses relatively close, trees area environment, primarily stories tall, environment relatively green
				Realistic Seattle aerial satellite top view image with high quality details with buildings and roads in Seattle that probably has the following objects and characteristics: cars parked lot, nearly parking lot, foreground cars parked, white building background, trees foreground cars

Figure 3. More **same-area** generated results from our proposed GPG2A. From left to right are the input ground image, target aerial image, predicted BEV layout, synthesized aerial image, and corresponding text prompt.

aerial image, and the same condition is used for all training samples. The constant text prompt was as follows: *”Realistic aerial satellite top view image with high-quality details with buildings and roads”*.

The city text prompt adds a dynamic variable that indicates which city the corresponding sample belongs to. We believe that this addition helps guide the model to differentiate between cities, thus, generating better and diverse samples. This prompt was defined as follows: *”Realistic {CITY} aerial satellite top view image with high-quality details with buildings and roads in {CITY}”*.

The raw text prompt directly conditions Gemini’s output descriptions. An example of such a prompt can be seen in Fig. 7. These descriptions carry excessive details about the ground image, leading to noise synthesis as shown in Tab. 4 in the main script. Thus, we developed our dynamic text which extracts only relevant information from the collected Gemini descriptions. This relevant information was represented in the form of key phrases. The dynamic text prompt utilized both dynamic city assignment and key-phrase extraction, and was structured as follows: *”Realistic CITY aerial satellite top view image with high-quality details with buildings and roads in CITY that probably has the following objects and characteristics: KEYWORDS”*. Examples

of dynamic texts are shown in Figs. 3 and 4.

To extract the key phrases from the raw Gemini output, we adopt the Maximal Marginal Relevance (MMR) technique [4]. MMR can be used to rank key phrases based on their relevance to the text content, while also considering their diversity. The ranking score M is given as follows,

$$M \stackrel{\text{def}}{=} \underset{D_i \in \mathcal{S}}{\text{argmax}} [\lambda(\Psi(D_i, Q)) - (1 - \lambda)\max_{D_j \in \mathcal{S}}(\Psi(D_i, D_j))], \quad (2)$$

where Q is the Gemini query text, D_i denotes the key phrase to be ranked, D_j are all other remaining key phrases in the document (excluding D_i), \mathcal{S} is the set of all ranked key phrases, $\Psi(\cdot, \cdot)$ stands for the cosine similarity between two phrases. This equation controls the trade-off between the relevance and diversity of the extracted key phrases by the λ parameter. In our experiments, we empirically set the diversity parameter $\lambda = 0.3$, $m = 5$, and $N = 3$, which yielded the best trade-off between diversity and relevance.

Different text prompts with the same layout: To qualitatively measure how different text prompts affect the synthesis, we visualize the generated aerial images of the same layout but varying their input text. Fig. 8 illustrates three comparisons between using our dynamic text prompt and a modified prompt. The first experiment was designed to

Ground Image	Target Aerial Image	GPG2A (ours)		
		BEV Layout	Aerial Image	Dynamic Text
				Realistic SanFrancisco aerial satellite top view image with high quality details with buildings and roads in SanFrancisco that probably has the following objects and characteristics: area houses crossroad, residential area houses, houses stories tall, street relatively narrow, trees bushes street
				Realistic SanFrancisco aerial satellite top view image with high quality details with buildings and roads in SanFrancisco that probably has the following objects and characteristics: park trails park, trails park covered, park covered green, trails background hill, green grass plants
				Realistic Chicago aerial satellite top view image with high quality details with buildings and roads in Chicago that probably has the following objects and characteristics: sidewalks street relatively, trees sidewalks street, long urban street, street relatively wide, street tall buildings
				Realistic Chicago aerial satellite top view image with high quality details with buildings and roads in Chicago that probably has the following objects and characteristics: cars right street, smaller street left, wide street foreground, building street sign, large brick building
				Realistic Chicago aerial satellite top view image with high quality details with buildings and roads in Chicago that probably has the following objects and characteristics: parallel streets park, trees sides streets, residential area parallel, streets lined apartment, apartment buildings trees

Figure 4. More **cross-area** generated results from our proposed GPG2A. From left to right are the input ground image, target aerial image, predicted BEV layout, synthesized aerial image, and corresponding text prompt.

Ground Image	Target Aerial Image	Generated		Dynamic Text
		BEV Layout	Aerial Image	
				Realistic Chicago aerial satellite top view image with high quality details with buildings and roads in Chicago that probably has the following objects and characteristics: residential area lowrise, fourway street intersection, intersection center buildings, mixed commercial buildings, sidewalks streets relatively
				Realistic Chicago aerial satellite top view image with high quality details with buildings and roads in Chicago that probably has the following objects and characteristics: sidewalk road surrounded, wide road sidewalk, trees sidewalk street, street relatively clean, variety buildings including
				Realistic NewYork aerial satellite top view image with high quality details with buildings and roads in NewYork that probably has the following objects and characteristics: area densely populated, walkway area densely, urban area tall, buildings large road, intersection foreground trees

Figure 5. Some randomly sampled failure cases generated by our GPG2A. From left to right are input ground images, target aerial images, predicted BEV layout maps, and generated aerial images.

test the effect of changing both the city and keywords of the prompt. The style of the image changed from Chicago to a

residential area in Seattle. The second sample only changes the city but fixes the keywords. The type of intersection

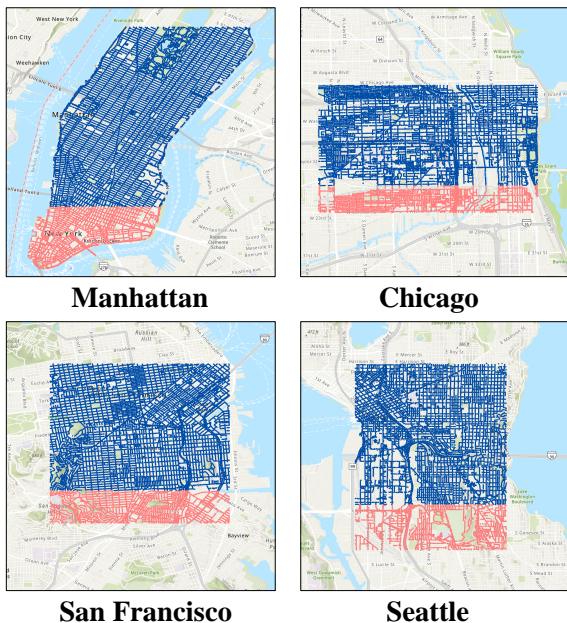


Figure 6. Visualization of training and testing splits for each city in our VIGORv2 dataset. Blue lines indicate the training portion. Red lines indicate the testing portion.

generated is different, representing New York streets. The last sample visualizes changing only the keywords but keeping the city as is, Chicago. The overall scenery and style of the generated images were similar due to sharing the city. But a highway-like street could be seen in the generated image, as the prompt had 'highway' in its keywords. These three visualizations underscore the importance of the text description in the synthesis.

10. Class Significance Visualization

The choice of classes in the BEV layout map was based on frequency, meaning that the most frequent classes in the dataset were used. To show the significance of specific classes, we visualize the generated images by omitting some of them. Three samples can be found in Fig. 9 where we first exclude the 'building' class, resulting in an image without any buildings. Similarly, the other two examples show images without parking spots and buildings when these classes were omitted.

11. GeoDTR+ augmentation

As mentioned in the manuscript, we also applied the data augmentation techniques in a more recent cross-view geolocalization model, GeoDTR+ [21]. The settings are the same as described in Sec. 6.1 of the manuscript. Table 3 summarizes the results. Similar to the conclusion in the manuscript, we also observed cross-area performance im-

provement on GeoDTR+ with our data augmentation. For example, R@1 increases from 63.65% to 65.71% on cross-area while $p_o = 0.6$. However, we did not observe the performance increase in the same-area protocol. This might be the better baseline performance of GeoDTR+ [21] than SAFA [15]. So that the data augmentation cannot significantly improve the same-area performance.

	p_o	R@1↑	R@5↑	R@10↑	R@1%↑
same-area	0	90.21%	96.39%	97.51%	99.47%
	0.4	90.05%	96.30%	97.48%	99.61%
	0.6	90.50%	96.46%	97.57%	99.54%
	0.8	90.11%	96.34%	97.33%	99.45%
cross-area	0	63.65%	81.01%	86.46%	96.85%
	0.4	65.48%	82.16%	87.39%	97.03%
	0.6	65.71%	82.53%	87.80%	97.13%
	0.8	64.54%	81.39%	87.15%	96.93%

Table 3. Results of our data augmentation on SAFA. $p_o = 0$ indicates no augmentation is applied.

12. Sketch-based Region Search Discussion

This section presents a more detailed discussion of the sketch-based region search application presented in Sec. 6.2 of the main manuscript. We show additional samples, including an intersection and a swirly road, as shown in Fig. 10. It is clear how the top retrieved images represent the given sketch and text. Although the street layout does not exactly match the third example, the environment is similar, a residential area in a warm climate. This discrepancy may result from the limited pool of referencing aerial images, where an exact match for the swirly road may not exist, but a close one does.

We observed consistent retrieval results even with a larger aerial database. For example, we tried retrieving aerial images from the four cities of VIGORv2, as shown in Fig. 11. Overall, the retrieved images align well with the given sketch and text in all cities. Any inconsistencies, if present, would appear in the layout rather than the environment. For instance, all images in the last row represent a highway, a parking lot exists in the second-row images, and an intersection appears to the right of the fourth-row images. These results demonstrate GPG2A's robustness and how it is applicable in cross-domain applications.

As mentioned in the main manuscript, we conducted a survey with 61 participants to quantitatively evaluate the retrieval results of our pipeline. The images shown to each volunteer are illustrated in Fig. 10. Participants rated how well the retrieved image corresponded to the given sketch and text pairs on a scale from 1 to 10. Ratings above 5 were considered agreements that the image matched the sketch and text pair, while ratings below 5 were considered disagreements. With this convention, 66%, 60%, and 24% of



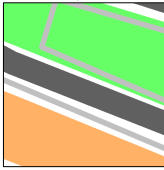


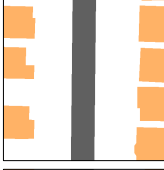




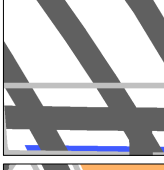
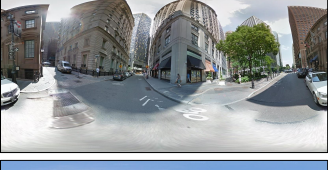

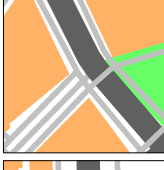



Ground	Aerial	Layout	Gemini Text
			This is a tree-lined block in an urban area. There are several large apartment buildings made of brick and a few cars parked on the street. There is a park in the center of the block with a lawn benches and a few trees.
			This is a residential area with about 7 houses on both sides of a wide road. The houses are mostly one or two stories tall and have a variety of architectural styles. The area is fairly green with many trees and lawns.
			This is a residential area with a few houses on both sides of the road. The houses are mostly two or three stories tall and have a variety of colors and styles. The road is wide and has a few cars parked on it. There are also a few trees and bushes on the side of the road.
			This is a wide elevated highway with a concrete barrier in the middle. The highway is surrounded by a green landscape with many trees. Tall buildings made of glass and concrete line the horizon in the distance.
			The image shows a wide street with several large buildings on both sides. The buildings are mostly made of stone or brick and have a lot of windows. There are a few trees on the street and some people walking around. The street is paved with cobblestones and there are a few cars parked on the side of the road.
			This is a residential area with low-rise apartment buildings on both sides of the road. There are a few trees on the sidewalks and the street is fairly wide with a clear view of the surroundings.

Figure 7. More samples from our proposed VIGORv2 dataset. From left to right are ground images, aerial images, BEV layout maps, and text descriptions for ground images.

the volunteers agreed that the top-1, top-5, and randomly selected images, respectively, represent the given sketch and text.

13. Limitations

In this paper, we paved the way for this challenging research, introducing a new dataset as well as new image quality measurements as a foundation for future comparative studies in this emerging field. The challenges of this task arise from significant variations in perspective, occlusions of objects, and the varying range of visibility between aerial and ground views. Despite these challenges, our proposed algorithm demonstrates proficiency in preserving geometric information due to the explicit conditioning of the layout maps into our model. One of the limitations of our

algorithm is the inability to generate images capturing the location of movable objects like cars and pedestrians. This limitation resulted from the unavailability of synchronized training data, particularly in obtaining a timely synchronized ground-aerial dataset. Looking ahead, one promising avenue for future exploration involves further enhancing our proposed models for the quality of image synthesis. Additionally, addressing the scarcity of synchronized datasets, future work could focus on the creation of larger and synchronized datasets including diverse cities across continents. This expansion aims to enable the scalability of the proposed methods, advancing more comprehensive and globally applicable systems.

Ground truth	Modified prompt		Original prompt	
	Generated image	Keywords	Generated image	Keywords
		Seattle, residential house, two-story house, home, playground, family residential building		Chicago, urban street, buildings, wide lanes, sidewalks, businesses
		New York, four-way street, intersection, streetlights, commercial buildings, sidewalks		Chicago, four-way street, intersection, streetlights, commercial buildings, sidewalks
		Chicago, Highway		Chicago, road, intersection, urban, brick buildings, sidewalk

Figure 8. Varying the text prompt greatly influences the generated image. Keywords such as highway, residential house, or city name are reflected in the generated image

Control class	Ground truth	w/ class	w/o class
Building			
Parking			
Road			

Figure 9. Visualization of omitting Building class. The generated image loses important details.

14. Societal Impact

In this paper, our novel GPG2A is effective in many areas, as stated in the main script, such as land use classification [2, 18], urban planning [16], destruction detection [1], transportation [5, 8, 12] and socioeconomic studies [3, 14]. The predicted BEV layout can also be an auxiliary signal to the positioning system, i.e. comparing the BEV layout to the map to estimate the location. Thus, to this end, the proposed GPG2A will advance the research in both cross-view image synthesis and image geo-localization which will eventually benefit the society. Our proposed VIGORv2 dataset is complementary to the original VIGOR dataset with newly collected center-aligned aerial images, BEV layout maps, and text descriptions for ground images. This

dataset will advance future research in this direction and inspire further researchers to work on this problem. Our new aerial image quality evaluation metrics provide a new tool in this topic to help researchers understand the quality of the synthesized images. To this end, this work will benefit the community and advance the research in this area.

References

- [1] Destruction from sky: Weakly supervised approach for destruction detection in satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162:115–124, 2020. 8
- [2] Abolfazl Abdollahi and Biswajeet Pradhan. Urban vegetation mapping from aerial imagery using explainable ai (xai). *Sensors*, 21(14):4738, 2021. 8
- [3] Jacob Levy Abitbol and Marton Karsai. Interpretable socioeconomic status inference from aerial imagery through urban patterns. *Nature Machine Intelligence*, 2(11):684–692, 2020. 8
- [4] 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. 4
- [5] 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. 8
- [6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1
- [7] 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. 1
- [8] Ruimin Ke, Zhibin Li, Jinjun Tang, Zewen Pan, and Yin-hai 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. 8
- [9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1
- [10] 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. 1
- [11] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 1

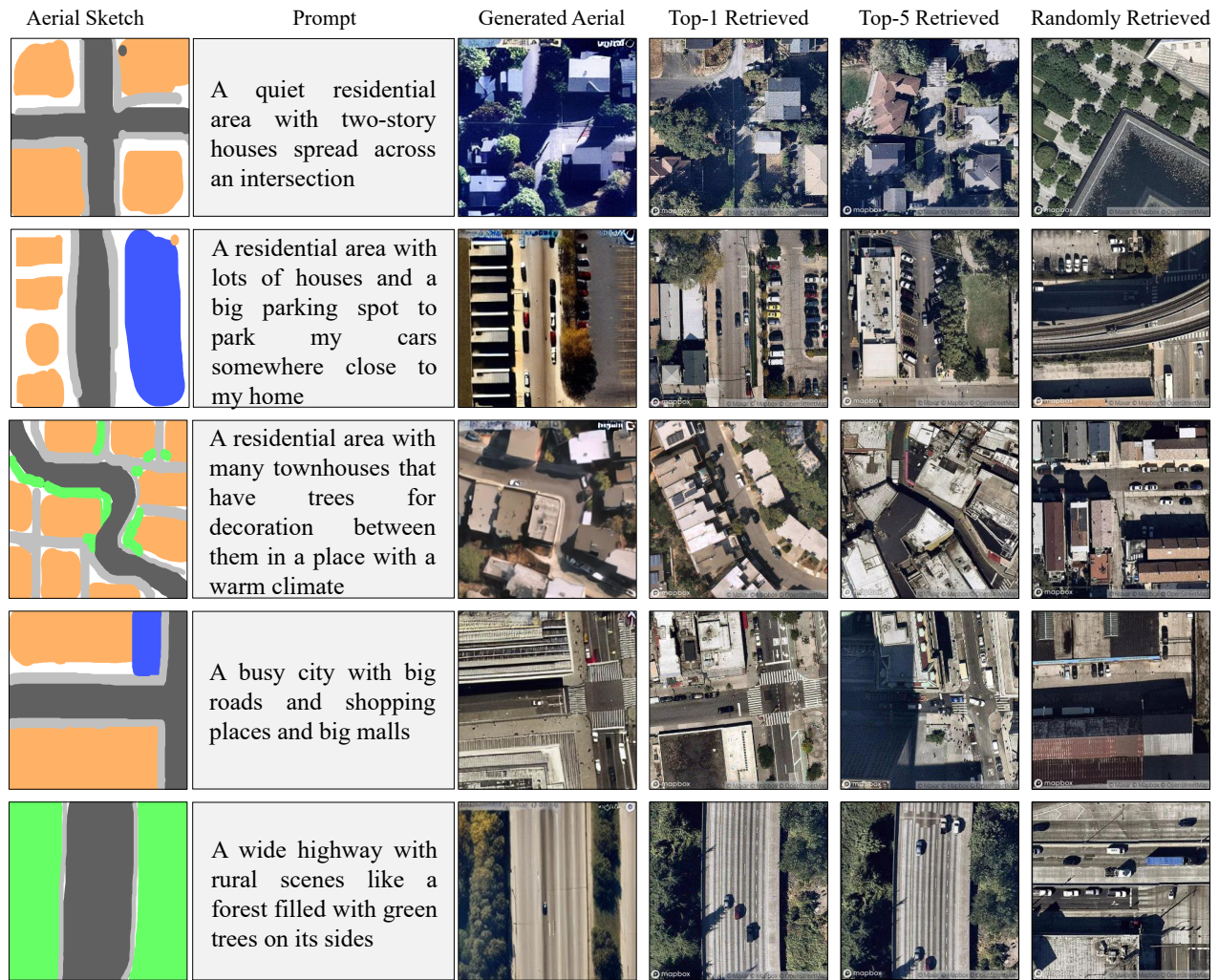


Figure 10. More retrieval samples, and the survey images shown to the participants. We analyzed five groups of images in three cases, top-1, top-5, and a randomly selected image

- [12] 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. 8
- [13] 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
- [14] 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. 8
- [15] Yujiao Shi, Liu Liu, Xin Yu, and Hongdong Li. Spatial-aware feature aggregation for image based cross-view geolocalization. *Advances in Neural Information Processing Systems*, 32, 2019. 1, 6
- [16] 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. 8
- [17] 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. 1, 2
- [18] 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. 8
- [19] 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
- [20] 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*

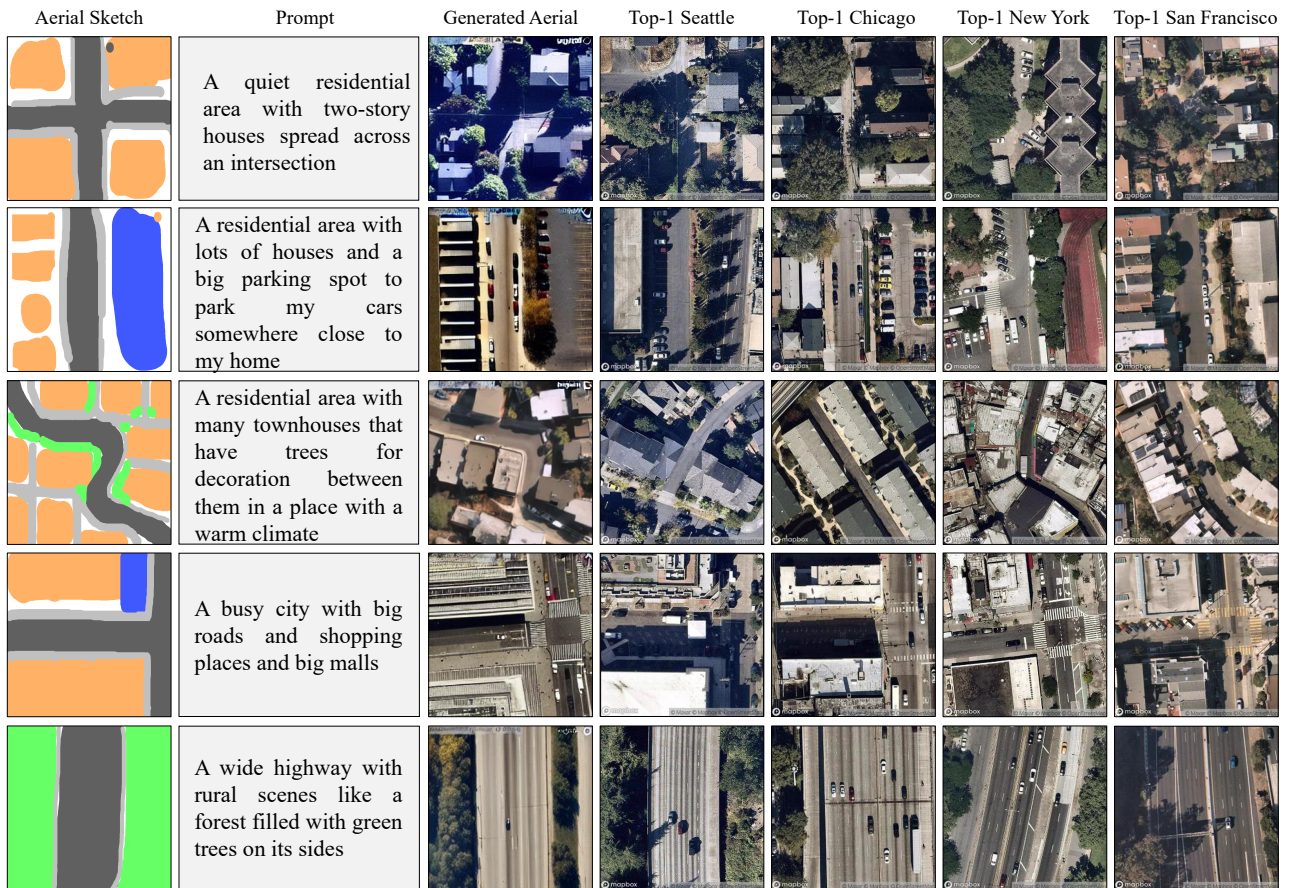


Figure 11. More cross-city results for the sketch-based region search application. Most images reflect both the sketch and text, even across different cities

IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 1, 2

- [21] 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. 6
- [22] Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Cross-view image sequence geo-localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2914–2923, 2023. 2