Cross-View Fusion for Occluded Roads Inpainting in Satellite Imagery

Afrina Tabassum*, Hesham M. Eraqi*, Yang He, George Forman, Mohamed Moustafa
Last Mile, Amazon, WA, USA
{afrinat, heraqi, yanhea, ghforman, mmoustm}@amazon.com

***Abstract*—Remote sensing imagery is a a rich source that serves a wide range of applications including urban planning, land management, environmental monitoring, and digital map enrichment by detecting roads and building outlines. Nevertheless, occlusions from trees often hide important features like roads from bird's-eye viewpoint. To tackle this problem, we propose GeoRoadInpaint model, leveraging stable diffusion to reveal roads hidden by tree occlusions. Additionally, we introduce GeoFusionRoadInpaint model, which extends GeoRoadInpaint by incorporating street-view imagery from a vehicle-mounted camera to guide more precise road generation that better matches real-world. Both quantitative and qualitative results demonstrate that the proposed approach successfully removes occlusions and generate realistic roads. Testing on a comprehensive dataset of 4,500 satellite images shows an 18.18% increase in TOPO score for detecting roads from the inpainted images compared to detecting them from raw imagery.**

***Index Terms*—Satellite image inpainting, roads extraction.**

I. Introduction

Digital maps are increasingly vital for autonomous driving, package delivery, fleet management, and ride-sharing. Accurate, up-to-date, feature-rich, and high-coverage maps are key for safe and efficient route planning and navigation to support those applications [1] [2]. Satellite and aerial imagery provide critical input for improving map accuracy, freshness, richness, and coverage. Computer vision models use these images to extract map features such as roads [3], road attributes [4], building outlines [5], and entrance locations [6]. These images also aid manual editing of map data [1] and driver navigation.

Road detection becomes difficult or impossible in areas with substantial tree occlusions. Our proposed models address this issue by enabling visibility through occlusions using diffusion inpainting processes. The proposed GeoRoadInpaint model inpaints the satellite image to show roads occluded by vegetation, given road masks created from vehicle GPS traces, which are readily available through navigation services. In addition, GeoFusionRoadInpaint enhances the process by incorporating street-view imagery alongside road masks to generate roads that more precisely match reality as described in Figure 1.

While we focus on removing tree occlusions, our approach can extend to other occlusion patterns such as high-rise buildings and shadows. Both quantitative and qualitative results confirm our models successfully remove occlusions and generate realistic roads, enabling more accurate road extraction from remote sensing imagery.

*Equal contribution

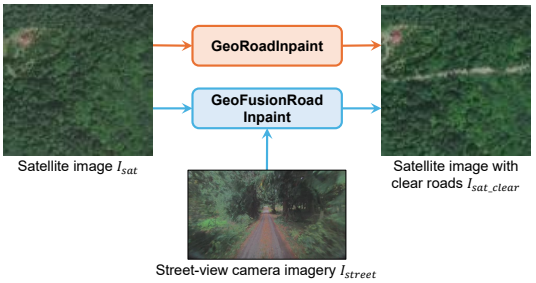

Fig. 1. Our two proposed models: 1) GeoRoadInpaint which inpaints a remote sensing image given the roads mask created from GPS traces of vehicles that traveled through the region and 2) GeoFusionRoadInpaint which additionally utilizes street-view imagery to generate more precise and realistic roads.

II. Related Work

Recent advances in diffusion models have revolutionized image generation across various domains, including remote sensing applications. While general-purpose image generation has progressed from text-to-image synthesis [7, 8] to sophisticated editing [9–11], inpainting [12, 13], and composition [14], these models are not tailored for the fine-grained precision required in satellite imagery applications.

In the remote sensing domain, Generative Adversarial Networks (GANs) have been widely adopted for super-resolution tasks [15–17]. More recently, diffusion models have enabled remote sensing generation under multimodal conditions including text, OpenStreetMap (OSM) control, sketch, and road mask [18–20]. GeoSynth [20] utilizes a ControlNet and a location-tailored CLIP encoder [21] to condition a pretrained latent diffusion model (LDM) with OSM control and location information. While recent works have successfully performed satellite image inpainting for temporal changes such as wildfire, flooding, and wind prediction at the level of entire input satellite images, none have focused on spatial inpainting of particular features targeting specific sections in the input satellite imagery based on multimodal conditions while leaving other areas unaltered. To address this gap, we propose GeoRoadInpaint and GeoFusionRoadInpaint, which are inpainting models specifically tailored for fine-grained spatial editing in satellite imagery.

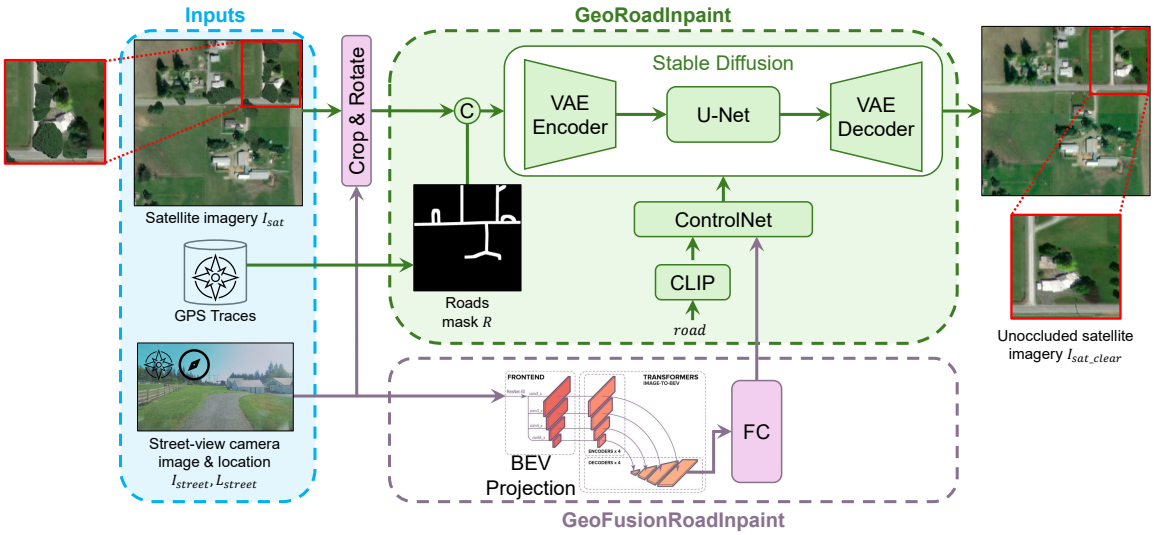

Fig. 2. The schematic architecture of the proposed models. The proposed GeoRoadInpaint model (green path in the figure) is based on the Stable Diffusion model receiving the satellite image I_{sat} having some ways occluded due to vegetation, along with roads mask R that is constructed from historical GPS traces of vehicles that traveled through the region, to inpaint the input satellite image I_{sat} to reveal the occluded roads in the output image I_{sat_clear}. The GeoFusionRoadInpaint variant leverages an additional bird's-eye-view (BEV) encoder and ControlNet to condition the generation process to generate roads similar to reality captured in the input street-view images I_{street}. The model is trained on a synthetic dataset we generated. During training, the input satellite image is set to I_{syn_sat} having synthetic vegetation added and the model learns to output the clear original satellite image I_{sat}.

III. Methodology

A. Problem Definition

The task of "seeing through occlusion" aims to remove occlusion in a given remote sensing image to reveal the underlying roads given external knowledge from vehicle GPS traces and street-view front-facing camera images. Formally, given a satellite image, I_{sat} where some roads are occluded, the objective is to inpaint I_{sat} and generate I_{sat_clear} where all the occluded roads are revealed.

To this end, we propose two variants of satellite image inpainting model GeoRoadInpaint and GeoFusionRoadInpaint. To our knowledge, GeoRoadInpaint is the first spatial inpainting model for geospatial data. It is an image inpainting model, with learnable parameters $\theta_1 : (I_{sat}, R) \rightarrow I_{sat\_clear}$ where R is a roads binary mask, corresponding to I_{sat}, set to a value of 1 at pixels associated with roads in I_{sat} and 0 otherwise. The roads mask R is constructed from historical GPS traces of vehicles that traveled through the region covered in the input I_{sat} image. On the other hand, GeoFusionRoadInpaint is an extension of GeoRoadInpaint which utilizes the street-view images to enhance the quality of the generated roads and to make them more akin to real roads by learning additional parameters $\theta_2 : (I_{sat}, R, I_{street}, L_{street}) \rightarrow I_{sat\_clear}$ where I_{street} is a street street-view image and and L_{street} is the associated vehicle GPS and IMU readings at the image capturing time.

B. Synthetic Dataset Generation

We formularize seeing through the occlusion as a supervised training problem. We generate a synthetic dataset of pairs of satellite images before and after tree occlusion removal, and use it to train the model. The training dataset is composed of satellite images with roads. In each image, we inject synthetic trees to the road parts which are not originally occluded by trees by performing the following steps.

1) Vegetation Mask: Firstly, we segment the trees and vegetation present in a satellite image using an image segmentation model based on textual prompt. We use the SAMGeo [22] model, which is an application of Segment Anything Model (SAM) [23]for geospatial data analysis. We perform the segmentation using the textual prompt "trees" to produce the segmentation binary mask. Figure 3 demonstrates the overall process on a sample satellite image I_{sat}, showing the negative of the vegetation mask segmented using SAMGeo $\bar{V}$.

2) Unoccluded Roads Mask: We utilized three months of GPS traces from vehicles that traveled through the regions covered our dataset to create roads mask images. For each satellite image I_{sat}, a corresponding roads mask R is constructed. GPS raster images are created from the traces following the method in [24]. Afterwards, the rasters are smoothed with a Gaussian filter and then binarized. Empirically, the Gaussian

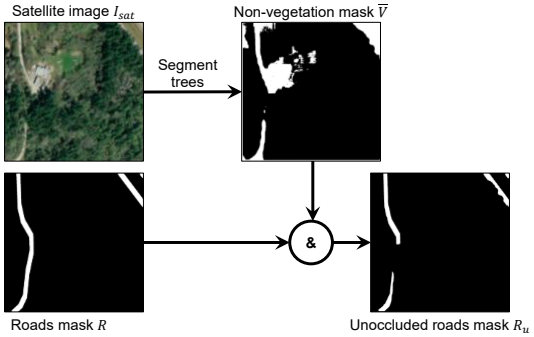

Fig. 3. Procedure of creating roads mask excluding regions occluded by vegetation.

Fig. 4. Examples of original (left) and created synthetic (right) remote sensing images.

filter standard deviation σ and the binarization threshold are set to 7.0 and 0.5 respectively. The binary image is dilated to produce the roads mask R. Finally, as illustrated in the example in Figure 3, the unoccluded roads mask R_u is the pixel-wise multiplication of the roads mask R and the non-vegetation $\bar{V}$.

C. Artificial Tree Patch Addition

In this step we utilize the SAM model [23] to cut out tree patches from several satellite images. We selected a set B of 40 diverse tree patches from several images to be used in our dataset. In each satellite image in the training dataset, we paste a random selection of tree batches from B based on the following criteria:

- Placement on unoccluded road areas
- Color matching with the original image's trees
- Diversified color distribution and occlusion percentage for better generalization

We refer to the generated synthetic satellite image as I_{syn_sat}. During training, the input satellite image is set to I_{syn_sat} having the synthetic vegetation added and the ground-truth is set to the original satellite image I_{sat}. Additionally, we paste some tree patches randomly outside the unoccluded roads region in R_u to both of I_{syn_sat} and I_{sat}, in order for the model to avoid learning to remove trees of a specific distribution in B. Figure 4 shows two examples of synthetic images created using the aforementioned method.

D. GeoRoadInpaint

GeoRoadInpaint is based on the stable diffusion (SD) model [8]. The model inpaints on a satellite image based on a roads mask, generating pixels near the occluded roads and preserving the original image elsewhere. The SD model takes the satellite image and the roads mask as input and generates inpainted image by adding roads that were occluded due to vegetation.

During training, I_{syn_sat} and R are passed as input to the SD model along with the text prompt: "ROAD" to the CLIP text encoder [25]. The variational autoencoder (VAE) [26] encodes the images into latent vectors $l_{I_{syn\_sat}}, l_R$. Then, during training, noise is added to the ground-truth I_{sat} and the U-NET architecture [27] is trained to predict the added noise with the guidance of the CLIP output embeddings. The Mean Squared Error (MSE) loss between the real and predicted noise is calculated to train the SD model. During inference, the original image and the corresponding roads mask are provided as input along with noise. Then, U-Net generates the edited image latent vectors from that noise. Then the generated latent vectors are passed to the VAE decoder to generate the image in the same shape of original image. Figure 2 depicts the overall model architecture.

E. GeoFusionRoadInpaint

While GeoRoadInpaint removes occlusion from the original image by generating realistic roads, there is no guidance provided on how the generated road should visually look like. Even though the generated roads look realistic, they don't resemble the actual road virtual features such as road color, texture, and lane markings. To this end, we propose GeoFusionRoadInpaint which extends GeoRoadInpaint to take additional guidance from the street-view images. We train GeoFusionRoadInpaint on top of GeoRoadInpaint by keeping the parameters of GeoRoadInpaint frozen.

For a street-view image I_{street} with associated vehicle GPS and IMU information L_{street} at capture time, the input satellite image I_{sat} is 1) cropped around the vehicle location such that vehicle position becomes at the center and 2) rotated such that the vehicle traveling direction becomes heading upwards in the satellite image. This standardizes the association between the input satellite image and the perspective captured in the street view image. Bird's eye view (BEV) features are extracted from the street-view image, and then we pass the control I_f into the image diffusion model.

1) BEV Projection: BEV features are aligned with satellite imagery point of view. Thus, we utilize the feature extraction transformers-based model introduced in [28] for translating street-view images into maps. As depicted in Figure 2, the projection network consists of a ResNet-50 [29] backbone which extracts spatial features in the image-plane at multiple scales to allow for scale invariance, encoder-decoder transformers to translate features from the image-plane to BEV perspective, and a linear fully connected (FC) layer to extract BEV features I_f. We use the pretrained weights for the ResNet-50 backbone and encoder-decoder transformers and train only the FC layer.

2) ControlNet: The extracted BEV features I_f from the street-view image are used to control image generation. For this purpose, we utilize ControlNet [30], which is a neural network architecture to add spatial conditioning controls to large pretrained text-to-image diffusion models. During training, we use the pretrained weights of our trained GeoRoadInpaint, and train only ControlNet and the BEV projection FC layer so the overall model learns to guide the image generation based on the BEV features control from the street-view imagery.

IV. Experiments

A. Dataset

We prepared $20,500$ pairs of satellite and street-view images, randomly split into $16,000$ model training and $4,500$

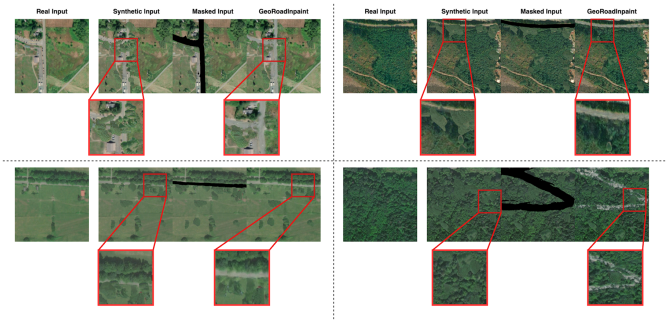

Fig. 5. Qualitative results from the GeoRoadInpaint model. Road masks are created from GPS traces of vehicles that traveled through the region.

TABLE I
QUANTITATIVE EVALUATION OF THE GEOROADINPAINT MODEL

Image Source	TOPO (↑)
Original	0.22
Finetuned-SD	0.20
Finetuned-LoRA-SD	0.21
GeoRoadInpaint	**0.26**

testing samples. Ground-truth road graphs are available for all satellite images recorded as lists of latitude and longitude points describing road geometries.

B. Evaluation Metric

To evaluate the proposed model, we use the TOPO score [31, 32]. We utilize DLinkNet [33] for roads extraction from the generated satellite images and compare the extracted roads with the associated ground-truth roads graphs to estimate the TOPO score. A higher TOPO score indicates the model is able recover more occluded roads in the input satellite image, which means the predicted roads are closer to the ground-truth reality. The TOPO score penalizes both cases when having missing roads (false negatives) and adding wrong extra roads (false positives) to the satellite images.

C. Baselines

For quantitative and qualitative analysis of GeoRoadInpaint, we utilize the following baselines:

Finetuned-SD: Finetuned the SD-inpainting model [8] with ground-truth satellite images I_{sat} without synthetic training data.

Finetuned-LoRA-SD: Finetuned the SD-inpainting model [8] using low rank adaptation (LoRA) [34] with ground truth satellite images I_{sat} without synthetic training data.

In addition, we provide a qualitative analysis of GeoFusionRoadInpaint generated images compared to the images generated by GeoRoadInpaint.

D. Evaluation

1) Quantitative Evaluation: Table I compares the TOPO scores for detecting roads from the improved satellite images generated by GeoRoadInpaint with respect to detecting them from the original satellite images and those generated by the baseline models. GeoRoadInpaint is demonstrated to improve roads detection accuracy by 18.18% compared to detecting roads directly from the raw satellite imagery before occlusion removal. These results, along with the qualitative analysis presented later in subsection IV-D2, demonstrate that the model is able to successfully unocclude roads in the input satellite images. Detecting roads from the satellite images produced by the Finetuned-SD and Finetuned-LoRA-SD baselines produces comparable TOPO performance that is similar to using the original satellite images. That indicates 1) the robustness of the LoRA method, which we followed during our models training, and 2) the effectiveness of the SD model to generate satellite images preserving the unoccluded roads present in the input original satellite images.

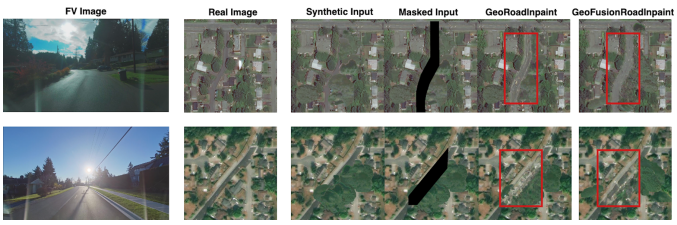

Fig. 6. Comparison between and the GeoRoadInpaint (top) and GeoFusion-RoadInpaint (bottom) models.

2) Qualitative Evaluation: Figure 5 shows more qualitative results of the model. The results demonstrate that GeoRoadInpaint successfully generated correct roads for the corresponding road masks from the vehicles' GPS traces by removing occlusion from trees for all the examples in the figures.

Figure 6 presents a comparison between GeoRoadInpaint and GeoFusionRoadInpaint on synthetic test samples. In the first row results in the figure, we observe that GeoFusionRoadInpaint generated roads which resemble the actual road seen in the street-view imagery having no lane markings, whereas GeoRoadInpaint generated roads that resemble other unoccluded roads present in the input satellite image. Similarly, in the second row results in the figure, GeoFusionRoadInpaint unrevealed the occluded road with a differentiated sidewalk seen in the street-view imagery, whereas GeoRoadInpaint generated roads of similar color without including the sidewalk marking.

V. CONCLUSION AND FUTURE WORK

We introduced GeoRoadInpaint, a satellite image inpainting model that generates roads the are occluded by vegetation. On a comprehensive test dataset of 4,500 real-world satellite images, both quantitative and qualitative evaluation demonstrated the proposed model successfully revealed occluded roads. It improved roads detection TOPO score by 18.18% compared to detecting roads directly from the input satellite imagery. As a secondary contribution, we proposed GeoFusionRoadInpaint which extends GeoRoadInpaint to incorporate street-view imagery for controlling the generation process. Qualitative evaluation suggests that traits of the generated roads are incorporated from the street-view images. For future work, we will further improve and analyze the GeoFusionRoadInpaint extension by leveraging more BEV encoding methods and extend the quantitative analysis.

REFERENCES

[1] V. Kango, H. M. Eraqi, and M. Moustafa, "High precision map conflation of fleet sourced traffic signs," in *IGARSS 2024 - 2024 IEEE International Geoscience and Remote Sensing Symposium*, 2024, pp. 4648–4648.

[2] G. Agarwal, "A principled approach to automated road network conflation," Ph.D. dissertation, University of British Columbia, 2021.

[3] Y. He, R. Garg, and A. R. Chowdhury, "Td-road: Top-down road network extraction with holistic graph construction," in *European Conference on Computer Vision*. Springer, 2022, pp. 562–577.

[4] Y. He, E. Eftelioglu, M. Moustafa, and A. R. Chowdhury, "A highly efficient and effective attribute learning framework for road graph from aerial imagery and gps," in *SIGSPATIAL BigSpatial workshop*, 2023, pp. 32–41.

[5] E. L. L. for Improved Building Polygon Extraction, "Can a computer frobnicate?" in *GeoIndustry workshop at ACM SIGSPATIAL*, 2025.

[6] G. Forman, "Getting your package to the right place: Supervised machine learning for geolocation," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2021, pp. 403–419.

[7] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[9] T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.

[10] C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon, "Sdedit: Guided image synthesis and editing with stochastic differential equations," *arXiv preprint arXiv:2108.01073*, 2021.

[11] B. Kawar, S. Zada, O. Lang, O. Tov, H. Chang, T. Dekel, I. Mosseri, and M. Irani, "Imagic: Text-based real image editing with diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 6007–6017.

[12] A. Lugmayr, M. Danelljan, A. Romero, F. Yu, R. Timofte, and L. Van Gool, "Repaint: Inpainting using denoising diffusion probabilistic models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 461–11 471.

[13] S. Xie, Z. Zhang, Z. Lin, T. Hinz, and K. Zhang, "Smartbrush: Text and shape guided object inpainting with diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 428–22 437.

[14] Y. Song, Z. Zhang, Z. Lin, S. Cohen, B. Price, J. Zhang, S. Y. Kim, and D. Aliaga, "Objectstitch: Object compositing with diffusion model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 310–18 319.

[15] P. Wolters, F. Bastani, and A. Kembhavi, "Zooming out on zooming in: Advancing super-resolution for remote sensing," 2023. [Online]. Available: https://arxiv.org/abs/2311.18082

[16] J. Cornebise, I. Oršolić, and F. Kalaitzis, "Open high-resolution satellite imagery: The worldstrat dataset–with application to super-resolution," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 979–25 991, 2022.

[17] Y. Gong, P. Liao, X. Zhang, L. Zhang, G. Chen, K. Zhu, X. Tan, and Z. Lv, "Enlighten-gan for super resolution reconstruction in mid-resolution remote sensing images," *Remote Sensing*, vol. 13, no. 6, p. 1104, 2021.

[18] S. Khanna, P. Liu, L. Zhou, C. Meng, R. Rombach, M. Burke, D. B. Lobell, and S. Ermon, "Diffusionsat: A generative foundation model for satellite imagery," in *The Twelfth International Conference on Learning Representations*, 2023.

[19] D. Tang, X. Cao, X. Hou, Z. Jiang, and D. Meng, "Crs-diff: Controllable generative remote sensing foundation model," *arXiv preprint arXiv:2403.11614*, 2024.

[20] S. Sastry, S. Khanal, A. Dhakal, and N. Jacobs, "Geosynth: Contextually-aware high-resolution satellite image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 460–470.

[21] K. Klemmer, E. Rolf, C. Robinson, L. Mackey, and M. Rußwurm, "Satclip: Global, general-purpose location embeddings with satellite imagery," *arXiv preprint arXiv:2311.17179*, 2023.

[22] Q. Wu and L. P. Osco, "samgeo: A python package for segmenting geospatial data with the segment anything model (sam)," *Journal of Open Source Software*, vol. 8, no. 89, p. 5663, 2023.

[23] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the International Conference on Computer Vision*, 2023, pp. 4015–4026.

[24] E. Eftelioglu, R. Garg, V. Kango, C. Gohil, and A. R. Chowdhury, "Ring-net: Road inference from gps trajectories using a deep segmentation network," in *Proceedings of the 10th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, 2022, pp. 17–26.

[25] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PmLR, 2021, pp. 8748–8763.

[26] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[27] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[28] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *2022 International conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 9200–9206.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.

[31] J. Biagioni and J. Eriksson, "Inferring road maps from global positioning system traces: Survey and comparative evaluation: Survey and comparative evaluation," *Transportation Research Record: Journal of the Transportation Research Board, 2291(1), 61-71*, 2012.

[32] P. Yin, K. Li, X. Cao, J. Yao, L. Liu, X. Bai, F. Zhou, and D. Meng, "Towards satellite image road graph extraction: A global-scale dataset and a novel method," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1527–1537.

[33] L. Zhou, C. Zhang, and M. Wu, "D-linknet: Linknet with pretrained encoder and dilated convolution for high resolution satellite imagery road extraction," in *CVPR workshops*, 2018.

[34] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.