TACKLING FEW-SHOT SEGMENTATION IN REMOTE SENSING VIA INPAINTING DIFFUSION MODEL

Steve Andreas Immanuel, Woojin Cho, Junhyuk Heo, Darongsae Kwon
TelePIX
07330, Seoul, South Korea
{steve, woojin, hjh1037, darong.kwon}@telepix.net

ABSTRACT

Limited data is a common problem in remote sensing due to the high cost of obtaining annotated samples. In the few-shot segmentation task, models are typically trained on base classes with abundant annotations and later adapted to novel classes with limited examples. However, this often necessitates specialized model architectures or complex training strategies. Instead, we propose a simple approach that leverages diffusion models to generate diverse variations of novel-class objects within a given scene, conditioned by the limited examples of the novel classes. By framing the problem as an image inpainting task, we synthesize plausible instances of novel classes under various environments, effectively increasing the number of samples for the novel classes and mitigating overfitting. The generated samples are then assessed using a cosine similarity metric to ensure semantic consistency with the novel classes. Additionally, we employ Segment Anything Model (SAM) to segment the generated samples and obtain precise annotations. By using high-quality synthetic data, we can directly fine-tune off-the-shelf segmentation models. Experimental results demonstrate that our method significantly enhances segmentation performance in low-data regimes, highlighting its potential for real-world remote sensing applications. All the codes are publicly available at https://github.com/SteveImmanuel/rs-paint.

1 INTRODUCTION

Remote sensing is a task to capture images of Earth's surface from a distance, typically using satellites. These data can then be utilized for many applications, such as climate forecasting (Troccoli, 2010; Palmer, 2014), marine ecosystem monitoring (Kavanaugh et al., 2021), urban planning (Malarvizhi et al., 2016). Due to the very high dimensional nature of satellite data, one of the most crucial part to process these data is to locate and segment any area of interest within these images. There has been a plethora of works in developing algorithm for object detection (Guo et al., 2018; Gong et al., 2022) and segmentation (Wu et al., 2019; Bahl et al., 2019; Karimov et al., 2024) to automate this process, notably using deep neural network.

However, like most neural networks, these models require extensive training data to achieve high performance. In the remote sensing domain, obtaining such datasets is particularly challenging. The images themselves are costly to acquire, often requiring access to specialized satellite systems or proprietary archives. Even when the data is available, privacy or security policies often restrict access to sensitive regions, e.g., military zones, further limiting usable training samples. Moreover, generating annotations is even more resource-intensive, as it requires domain expertise for accurate labeling, such as identifying environmental patterns, urban structures, or marine ecosystems. This combination of high costs and labor-intensive annotation processes significantly limits the availability of large-scale labeled datasets in the remote sensing domain.

Several works (Liu et al., 2023; Yang et al., 2023b; Hajimiri et al., 2023) have focused on developing few-shot learning algorithms for semantic segmentation to enable models to perform well with limited data. While these methods offer promising results, they usually only work on specific settings or require complex training strategy. We argue that a more effective approach to address this issue is to circumvent the data scarcity problem altogether. To this end, we propose leveraging inpainting diffusion models to generate additional training samples including the annotations at min-

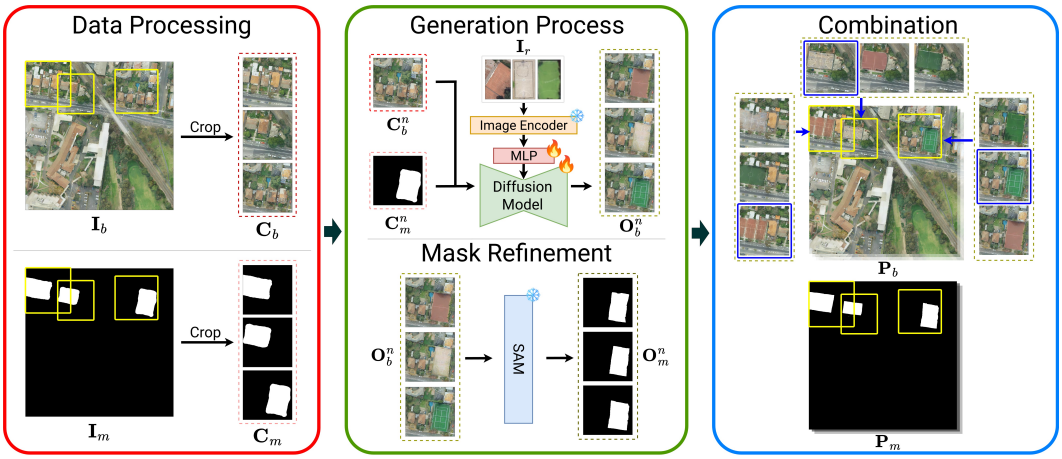

Figure 1: Overall pipeline of the proposed approach. An inpainting diffusion model generates novel-class samples, and SAM refines the segmentation masks. The results are used for training samples to improve model performance on few-shot settings.

imal cost. By conditioning the diffusion model on the object of interest, we can generate a diverse set of synthetic images featuring that object across a variety of scenes. Subsequently, we employ SAM to automatically derive the corresponding semantic masks. Although other data augmentation techniques, such as Copy-Paste (Ghiasi et al., 2021), have been explored, they have significant drawbacks. Specifically, while Copy-Paste can also increase the number of training samples, it often produces unrealistic images with noticeable artifacts along the object boundaries. When models are trained on such data, they may learn to rely on these artifacts as shortcuts for object detection or segmentation. Since these artifacts are absent in real-world images, the models trained in this manner tend to suffer from poor generalization performance.

2 Preliminaries

Few-shot segmentation. In few-shot segmentation setting (Tian et al., 2022), there are base classes and novel classes. The model is trained on abundant samples from the base classes and then adapted to segment instances of novel classes. Each novel class has a support set with a few annotated examples and a query set consisting of images to be segmented.

Image inpainting. Given a base image $\mathbf{I}_b \in \mathbb{R}^{H\times W\times 3}$, where H and W denote the height and width of the $\mathbf{I}_b$, and a binary mask $\mathbf{I}_m \in \{0,1\}^{H\times W}$, with 0 and 1 indicating unmasked and masked areas, respectively, the goal of image inpainting task is to fill the masked area with pixels that are semantically and structurally coherent with the unmasked areas. In our approach, we leverage image inpainting model to generate the novel classes within the masked area. Therefore, we essentially increase the number of annotated samples by generating more variations of the novel classes.

Image-conditioned diffusion model. Text-conditioned generative models, *e.g.*, Stable Diffusion (Rombach et al., 2022), have demonstrated impressive performance in generating realistic images. However, text prompts can be ambiguous, especially in the remote sensing domain as objects often appear similar from a top-down perspective and are difficult to describe solely with text. To address this, Yang et al. (2023a) and Song et al. (2023) use a reference image $\mathbf{I}_r \in \mathbb{R}^{H'\times W'\times 3}$ instead of a text prompt to condition the generation process while maintaining the object identity in the reference image. Following Yang et al. (2023a), $\mathbf{I}_r$ is processed through a frozen image encoder and compressed into a one-dimensional vector using multilayer perceptrons (MLP). This introduces an information bottleneck which enforces the model to learn the semantic information of $\mathbf{I}_r$ without collapsing into a trivial solution of copy-pasting $\mathbf{I}_r$ into $\mathbf{I}_b$.

3 Methodology

Given the few examples in the support set of novel classes in a segmentation dataset, we use Stable Diffusion to generate plausible variations of the novel classes in many different environments. The

generated samples help train an off-the-shelf segmentation model, mitigating overfitting that occurs when training with only the support set. This approach eliminates the need for a specially designed few-shot segmentation model and simplifies training by avoiding the typical two-phase process (Tian et al., 2022; Liu et al., 2023; Hajimiri et al., 2023), where the model is first trained on base classes and then adapted to novel classes.

3.1 Self-Supervised Training

To train Stable Diffusion with an image prompt, we need to collect pairs of $(\mathbf{I}_b, \mathbf{I}_r, \mathbf{I}_m)$ and the expected painted image $\mathbf{P}_b$. However, there are no publicly available datasets and it is infeasible to manually curate such dataset. Therefore, we leverage remote sensing object detection dataset for training in self-supervised manner. Given an image with a bounding box of an object in the image, we use the bounding box as the binary mask $\mathbf{I}_m$, the patch inside the bounding box as $\mathbf{I}_r$, and the original image as $\mathbf{P}_b$.

3.2 Generation Process

The overall generation pipeline is shown in Figure 1. Given a base image $\mathbf{I}_b$ and its mask $\mathbf{I}_m$, there are N plausible regions where an object can be generated. We first crop $\mathbf{I}_b$ into regions $\mathbf{C}_b = \{\mathbf{C}_b^n\}_{n=1}^N$ and $\mathbf{I}_m$ into their corresponding masks $\mathbf{C}_m = \{\mathbf{C}_m^n\}_{n=1}^N$. For each region $\mathbf{C}_b^n$, we independently generate L different variations $\mathbf{O}_b^n = \{\mathbf{O}_b^{n,l}\}_{l=1}^L$ using K different reference images $\{\mathbf{I}_r^k\}_{k=1}^K$. Due to the stochastic nature of Stable Diffusion, we can generate an arbitrary number of results even from a single reference image.

We find that the generated results are most realistic when the object mask covers approximately 15–30% of the cropped region. To ensure high-quality synthesis, we compute the cosine similarity between the generated object and its corresponding reference image using a CLIP encoder. If the similarity score falls below a predetermined threshold, we regenerate the sample to balance semantic consistency with diversity. Each variation generated for different regions is then combined to form complete augmented images $\mathbf{P}_b$ and their masks $\mathbf{P}_m$. Given a single base image, the total number of unique variations that can be generated is $\sum_{k=1}^N \binom{N}{k} L^k$, where $\binom{N}{k}$ accounts for the selection of k regions, and L^k accounts for the variations generated per selected region. The full algorithm and derivation are provided in Appendix C.

3.3 Mask Refinement

The masks $\mathbf{C}_m^n$ act as guidance for the model on where to paint the novel class object. However, the generated object may be smaller than the mask, leading to inaccuracies. To obtain a more precise segmentation mask $\mathbf{O}_m^n$, we leverage SAM (Kirillov et al., 2023), which has demonstrated strong zero-shot segmentation performance. While SAM does not inherently recognize object classes, this is not an issue in our case, as the generated object is conditioned on a reference image of a specific class, which effectively constraints the generated object class.

4 Experiments

4.1 Dataset

We use the few-shot set of the OpenEarthMap dataset (Broni-Bediako et al., 2024b) and select four novel classes for benchmarking: *boat*, *agriculture land*, *bridge*, and *sportsfield*. This dataset was also used for the OpenEarthMap few-shot challenge (Broni-Bediako et al., 2024a). Each class has only 5 annotated samples. For our approach, we initialize the Stable Diffusion model from the pre-trained checkpoint provided by Yang et al. (2023a). However, we replace the image encoder with pre-trained RemoteCLIP (Liu et al., 2024) and fine-tune the entire model on the SAMRS dataset (Wang et al., 2024a) for 100 epochs. This model is then used to generate additional samples for the novel classes. Specifically, for each class, we generate approximately 1,000 new samples using the annotated samples as the image conditioning. The generated samples can be found in Appendix A. Additional dataset details can be found in Appendix B.1.

Table 1: IoU comparison of various methods on different object classes. Results for challenge-winning methods are also included for reference. The underline indicates best results in each group and the **boldface** indicates best results overall.

Method	Class			
	Boat	**Agriculture Land**	**Bridge**	**Sports Field**
YOLOv11 (Jocher et al., 2023)	5.40	9.37	0.00	6.31
+ Copy-Paste (Ghiasi et al., 2021)	12.95	24.73	0.00	24.85
+ Ours	47.39	**46.35**	45.96	45.92
SegFormer (Xie et al., 2021)	6.00	26.49	5.65	29.27
+ Copy-Paste (Ghiasi et al., 2021)	13.18	30.99	4.94	29.08
+ Ours	45.02	37.46	23.27	45.10
Mask2Former (Cheng et al., 2022)	10.63	9.26	22.06	20.86
+ Copy-Paste (Ghiasi et al., 2021)	16.50	36.19	14.63	41.70
+ Ours	**72.53**	45.36	**57.24**	**66.36**
Challenge Winners				
SegLand (Li et al., 2024)	10.76	8.22	57.06	55.87
ClassTrans (Wang et al., 2024b)	0.00	44.78	6.94	49.98
FoMA (Gao et al., 2024)	58.64	1.44	17.01	40.59
P-SegGPT (Immanuel & Sinulingga, 2024)	0.00	32.36	0.00	38.50
DKA (Tong et al., 2024)	0.00	28.33	0.00	29.19

4.2 Performance Comparison

As baselines, we choose YOLOv11 (Jocher et al., 2023), SegFormer (Xie et al., 2021), and Mask2Former (Cheng et al., 2022) to represent different architectures. Each class is treated as a binary segmentation task, with a separate model trained for each class. We evaluate three training strategies: (i) Vanilla, using only the five annotated samples; (ii) Copy-Paste (Ghiasi et al., 2021), augmenting the five samples with copy-paste augmentation; and (iii) Ours, incorporating both the annotated and generated samples. More training details can be found in Appendix B.2.

The results are presented in Table 1. Models trained with only five annotated samples exhibit limited performance, often struggling with overfitting due to the scarcity of training data. Introducing the Copy-Paste augmentation improves performance in some cases, but it can also degrade results, as seen in the bridge class. This decline in performance is likely due to the inherent complexity of bridge objects, which need to connect two road segments in a structurally coherent manner. Simply copy-pasting bridge instances into new scenes does not guarantee a realistic connection between roads, potentially confusing the model and leading to suboptimal segmentation.

In contrast, our approach consistently delivers substantial improvements across all object classes and model architectures. By generating realistic novel-class samples through image inpainting, our method ensures that objects are naturally integrated into diverse environments, avoiding the pitfalls of naive augmentation techniques. This robustness underscores the flexibility and adaptability of our method, demonstrating its effectiveness regardless of the underlying segmentation model. Furthermore, we observe that models trained using our approach outperform can even outperform the challenge-winning submissions of the challenge. Importantly, our method achieves these results without relying on specialized architectures or complex training strategies, highlighting its practicality and ease of integration into existing workflows.

5 Conclusion

In this work, we introduce a simple yet highly effective approach for handling few-shot setting in segmentation tasks using an inpainting diffusion model. While we only conduct experiments for few-shot segmentation for remote sensing dataset, it is straightforward to adopt our approach for other task, such as object detection, as well as other domains, such as medical imaging and autonomous driving, where annotated data is scarce.

REFERENCES

Gaétan Bahl, Lionel Daniel, Matthieu Moretti, and Florent Lafarge. Low-power neural networks for semantic segmentation of satellite images. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*, 2019.

Clifford Broni-Bediako, Junshi Xia, Jian Song, Hongruixuan Chen, Mennatullah Siam, and Naoto Yokoya. Generalized few-shot semantic segmentation in remote sensing: Challenge and benchmark. *IEEE Geoscience and Remote Sensing Letters*, 2024a.

Clifford Broni-Bediako, Junshi Xia, Jian Song, Hongruixuan Chen, and Naoto Yokoya. Openearthmap land cover mapping few-shot learning challenge, May 2024b. URL https://doi.org/10.5281/zenodo.11396874.

Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 1290–1299, 2022.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3213–3223, 2016.

Tianyi Gao, Wei Ao, Xing-Ao Wang, Yuanhao Zhao, Ping Ma, Mengjie Xie, Hang Fu, Jinchang Ren, and Zhi Gao. Enrich distill and fuse: Generalized few-shot semantic segmentation in remote sensing leveraging foundation model's assistance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2771–2780, 2024.

Golnaz Ghiasi, Yin Cui, Aravind Srinivas, Rui Qian, Tsung-Yi Lin, Ekin D Cubuk, Quoc V Le, and Barret Zoph. Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2918–2928, 2021.

Hang Gong, Tingkui Mu, Qiuxia Li, Haishan Dai, Chunlai Li, Zhiping He, Wenjing Wang, Feng Han, Abudusalamu Tuniyazi, Haoyang Li, et al. Swin-transformer-enabled yolov5 with attention mechanism for small object detection on satellite images. *Remote Sensing*, 14(12):2861, 2022.

Wei Guo, Wen Yang, Haijian Zhang, and Guang Hua. Geospatial object detection in high resolution satellite images based on multi-scale convolutional neural network. *Remote Sensing*, 10(1):131, 2018.

Sina Hajimiri, Malik Boudiaf, Ismail Ben Ayed, and Jose Dolz. A strong baseline for generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11269–11278, 2023.

Steve Andreas Immanuel and Hagai Raja Sinulingga. Learnable prompt for few-shot semantic segmentation in remote sensing domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2755–2761, 2024.

Glenn Jocher, Jing Qiu, and Ayush Chaurasia. Ultralytics YOLO, January 2023. URL https://github.com/ultralytics/ultralytics.

Sardor Karimov, Dildora Sotvoldiyeva, Durbek Khalilov, and Nurillo Mamadaliyev. Deep neural network for semantic segmentation of satellite images. In *E3S Web of Conferences*, volume 587, pp. 03006. EDP Sciences, 2024.

Maria T Kavanaugh, Tom Bell, Dylan Catlett, Megan A Cimino, Scott C Doney, Willem Klajbor, Monique Messié, Enrique Montes, Frank E Muller-Karger, Daniel Otis, et al. Satellite remote sensing and the marine biodiversity observation network. *Oceanography*, 34(2):62–79, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Zhuohong Li, Fangxiao Lu, Jiaqi Zou, Lei Hu, and Hongyan Zhang. Generalized few-shot meets remote sensing: Discovering novel classes in land cover mapping via hybrid semantic segmentation framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2744–2754, 2024.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

Sun-Ao Liu, Yiheng Zhang, Zhaofan Qiu, Hongtao Xie, Yongdong Zhang, and Ting Yao. Learning orthogonal prototypes for generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11319–11328, 2023.

K Malarvizhi, S Vasantha Kumar, and P Porchelvan. Use of high resolution google earth satellite imagery in landuse map preparation for urban related applications. *Procedia Technology*, 24: 1835–1842, 2016.

Tim Palmer. Climate forecasting: Build high-resolution global climate models. *Nature*, 515(7527): 338–339, 2014.

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*, pp. 10684–10695, 2022.

Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18310–18319, 2023.

Zhuotao Tian, Xin Lai, Li Jiang, Shu Liu, Michelle Shu, Hengshuang Zhao, and Jiaya Jia. Generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11563–11572, 2022.

Jintao Tong, Haichen Zhou, Yicong Liu, Yiman Hu, and Yixiong Zou. Dynamic knowledge adapter with probabilistic calibration for generalized few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2781–2790, 2024.

Alberto Troccoli. Seasonal climate forecasting. *Meteorological Applications*, 17(3):251–268, 2010.

Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36, 2024a.

Shihong Wang, Ruixun Liu, Kaiyu Li, Jiawei Jiang, and Xiangyong Cao. Class similarity transition: Decoupling class similarities and imbalance from generalized few-shot segmentation. *arXiv preprint arXiv:2404.05111*, 2024b.

Ming Wu, Chuang Zhang, Jiaming Liu, Lichen Zhou, and Xiaoqi Li. Towards accurate high resolution satellite image semantic segmentation. *Ieee Access*, 7:55609–55619, 2019.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18381–18391, 2023a.

Yong Yang, Qiong Chen, Yuan Feng, and Tianlin Huang. Mianet: Aggregating unbiased instance and general information for few-shot semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7131–7140, 2023b.

A GENERATION SAMPLES

Figure 2: Comparison of inpainting methods for remote sensing: masked image (left), copy-paste (middle), and our method (right), showcasing realistic painted images for boats, agricultural land, bridges, and sportsfields.

B Additional Experiment Details

B.1 Dataset

The original few-shot set of the OpenEarthMap dataset comprises 408 samples, which are split into 258, 50, and 100 samples for training, validation, and test set, respectively. There are 7 base classes in the training set, 4 novel classes in the validation set, and another 4 novel classes in the test set. For our experiments, we select four of the eight novel classes to represent objects with varying complexity and scale. To generate the new samples, for each class, we randomly select 10 images from the training set to act as $\mathbf{I}_b$ and use the support set as $\mathbf{I}_r$. From these 10 images, we generate approximately 1000 unique variations, which are then used to train the segmentation models.

B.2 Training Details

During the fine-tuning of Stable Diffusion on the SAMRS dataset, the parameters of the image encoder used for conditioning are kept frozen, while the MLP and Stable Diffusion model remain trainable.

In the experiments, we use the following configurations for the baseline models:

- YOLOv11, X variant, pretrained on COCO dataset (Lin et al., 2014)
- SegFormer, B5 variant, pretrained on Cityscapes dataset (Cordts et al., 2016)
- Mask2Former, Large variant, pretrained on Cityscapes dataset (Cordts et al., 2016)

Using the corresponding training strategies, we train each model for 40 epochs, batch size of 32, and learning rate of 5e-5. We also use augmentations such as random cropping, random horizontal and vertical flipping, random rotation, random brightness scaling, and random gaussian blur.

C Combining Variations

Given the generation results for N different regions $\{\mathbf{C}_p^n\}_{n=1}^N$, where each region has L variations, we aim to generate all possible combinations of these regions. Each combination corresponds to a different variation of the final image. The idea is to combine selected regions from all possible subsets of the N regions, while considering all possible variations within each region.

The formalized steps are as follows:

1. **Binary Mask Representation**: To form the combinations, we generate all possible binary masks of length N, where the n-th bit corresponds to $\mathbf{C}_p^n$. A bit value of 1 indicates that the corresponding region is included, and a bit value of 0 indicates that the region is excluded.
2. **Combining Regions**: For each binary mask, we generate a variation by:
   - Including only those regions $\mathbf{C}_p^n$ for which the corresponding bit in the mask is 1.
   - For each included region, we select one of its L variations. Therefore, the number of combinations for a specific binary mask is L^k, where k is the number of selected regions, *i.e.*, the number of 1's in the binary mask.

The total number of unique variations can be computed by summing the possible combinations for all subsets of regions. Specifically, for any subset of size k, there are $\binom{N}{k}$ ways to select k regions, and for each of these subsets, there are L^k possible variations. Therefore, the total number of unique variations is $\sum_{k=1}^{N} \binom{N}{k} L^k$.

D Limitations

Currently the mask M needs to be manually created to ensure plausible location of the painted object. A promising direction for future work is to automate this process by leveraging a model to predict optimal object placement.

Additionally, the Stable Diffusion model does not account for the scale of objects relative to their surroundings. As shown in Figure 3, when a large mask area is used, the generated boat expands

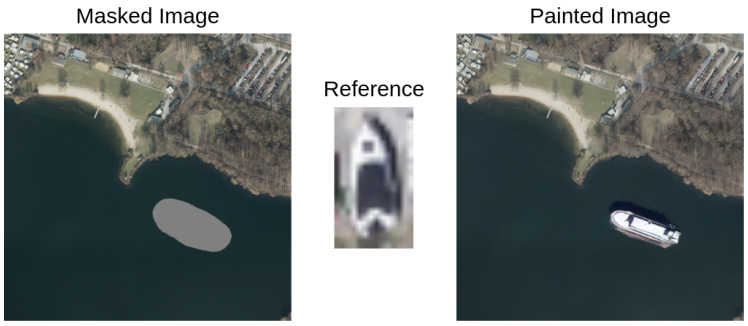

Figure 3: Failure case where the generated object is excessively large due to the large mask area.

to fill most of the space. However, when compared to surrounding objects such as buildings, it is noticeably oversized. A straightforward idea to tackle this issue is to incorporate the object scale information to condition the generation, which we leave for future work.