FLIP-TD: Free Lunch Inpainting on Top-Down Images for Robotic Tasks

Anukriti Singh*, Vishnu Sharma*, and Pratap Tokekar
Project Webpage: https://raaslab.org/projects/FLIP-TD

***Abstract*— Robotic systems are often limited by their sensor Field of View (FoV), which makes collision-free navigation and exploration in an unknown environment challenging. In contrast, humans are better at it because they can use their prior knowledge and consider the information beyond the FoV. What if robots could do it too? In our proposed approach, we aim to enhance the intelligence of robots by utilizing pre-trained masked autoencoders to make predictions of expanded FoV and synthesis a novel view. This allows the robot to reason and make informed decisions for safe and efficient navigation in unknown environments. We demonstrate the effectiveness of computer vision algorithms, specifically masked autoencoders, in solving practical robotics problems without the need for fine-tuning by using only top-down images. Our approach is evaluated in both indoor and outdoor environments, showcasing its performance in various settings of RGB, semantic segmentation, and binary images.**

I. INTRODUCTION

Mobile robot navigation through unknown areas has been studied by the robotics community for a long time. In the existing approaches, the robot updates the map based on its observations so far and moves according to the task at hand such as point-goal navigation, object navigation, and exploration. In the case of ground robots, the map is typically represented in a top-down view or Bird's Eye View (BEV), as the robot motion is constrained on the ground plane. Top-down representations are also used for aerial robots deployed for surveying and scouting.

The traditional approaches use the robot's observations with methods developed for the reliable construction of maps. Recent works across the wider robotics community have started exploring learning-based approaches to augment the robot's information to accomplish tasks. These methods help in modeling complex representations and processes and allow predicting future observations to improve robot safety and execution efficiency.

Learning-based methods face challenges in robotics due to the requirement for extensive datasets, which are typically sparser compared to computer vision applications. Simulators can generate virtual data, but introduce a ”sim2real gap” issue. Computer vision methods trained on large datasets may not directly apply to robot applications due to distribution differences in image representation. Fine-tuning is often used but requires similarity between pre-training and fine-tuning tasks, which can be challenging for robot navigation representations such as top-down images, semantic maps, and occupancy maps.

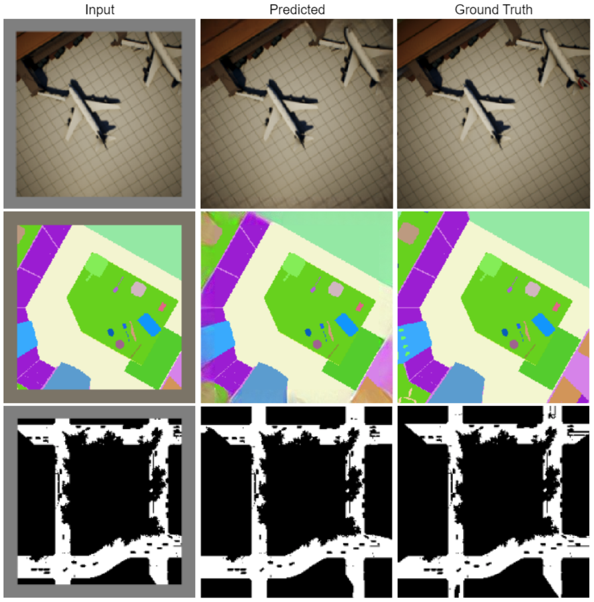

Fig. 1: Masked Autoencoder can be used to predict larger field-of-view in top-down images for RGB, semantic, and binary images without fine-tuning.

The recent emergence of foundational models, i.e. self-supervised models trained on huge datasets, aims to address this by including a variety of distribution of datasets. But how can we use these large self-supervised models in visual representation come at use for robotics tasks? In this paper, we study Masked Autoencoder (MAE) [12] for helping with predictions for top-down images.

A network capable of predicting, for example, the corners in an indoor environment, or the roads occluded by trees outdoors, can significantly help the robotics agent navigate the real world by utilizing the information beyond its field-of-view (FoV) for long-term planning. But it is important to note that the MAE is trained on front view/first person view images. And for robotics tasks, we need top-down, so to make predictions in top-down view images, should we fine-tune it? Surprisingly, the answer is no (for the most part). We test the effectiveness of MAE to synthesize novel views over indoor and outdoor images across three input modalities: RGB images, semantic maps, and binary maps.

*Both the authors contributed equally

This work is supported by the National Science Foundation under Grant No. 1943368 and ONR under grant number N00014-18-1-2829.

The authors are with the Dept. of Computer Science, University of Maryland, College Park, MD, USA {anukriti, vishnuds, tokekar}@umd.edu

We present scenarios where these predictions could improve task efficiency. Importantly, *we use only pre-trained MAE for all the experiments*, making MAE virtually a free-lunch method to help with a variety of tasks.

Specifically, we make the following contributions in this paper:

- We study MAE as an inpainting network for top-down images across RGB, semantic maps, and binary maps, and present quantitative and qualitative results across various degrees of increasing field-of-view.
- We present methods for semantics-guided inpainting with MAE in top-down images to remove the undesired objects for data enhancement.
- We show motivating examples to extract uncertainty in predictions from MAE that are helpful for uncertainty-guided navigation and exploration.

Our work highlights how MAE, and similar foundational models, can be used for robot tasks by choosing appropriate modalities without any fine-tuning, and paves the way for further improvement to the existing capabilities by task-specific tuning of these models.

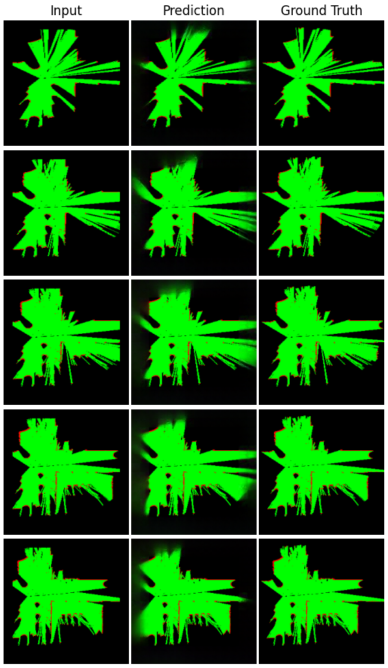

Fig. 2: Results on occupancy map from a TurtleBot2 robot.

II. Related Works

A. Top-Down Images for Robotics

Top-down images and map representations are vital for robot navigation and exploration. Navigating through an unknown map by Simultaneous Mapping and Localization (SLAM), which utilizes the robot's past observations, has been a cornerstone of robotics for robotics. A top-down semantic map is another representation of interest for robotic applications. These maps are useful for semantic goal navigation [10], [11]. Top-down images are also beneficial for tasks involving aerial robots such as surveying and scouting [4], [18]. The maps obtained by the aerial robots can be used for helping the ground robots navigate. Semantic maps are obtained from such images to identify navigable and non-navigable areas for the ground robot.

Recent works in this domain have sought to improve task efficiency by *predicting* the unobserved regions of the map to plan ahead. 2D Occupancy map, a top-down representation, has been the focus of many of these works, showing improvement in navigation, exploration distance, and time [14], [19], [24], [22]. Katyal et al. [15] show these benefits for high-speed navigation, highlighting the importance of predictions. While the predictions are limited to the perception module, planning can also enhance planning by extracting uncertainty from the predictions [13], [9]. The idea of uncertainty extraction also proves helpful in heterogeneous robot teams for risk-aware planning [23]. The key challenge with all these systems is that they need to be trained on the appropriate modalities, for which sufficient data may not be available, leading us to ask if there exist pre-trained models that can be used in these applications without much training effort, or better, without any fine-tuning at all?

B. Masked Autoencoders (MAE)

The idea of map prediction is similar to the well-known computer vision task of image inpainting [8]. As most of the existing networks are trained on first-person view RGB images, they do not work well on top-down images and other representations. Similar existing works for robotic applications rely on generative models [16], [20], [21], which require training or fine-tuning networks on simulation data to get accurate results. We instead turn our attention to self-supervised transformer-based networks that are capable to reason about shapes due to their ability to capture long-range relationships.

One such powerful network is masked autoencoder (MAE) that uses ViT encoder [7] and is trained to use only the visible patches of an image to predict the missing patches, similar to the training strategy of BERT [6]. MAE uses linear projections and position encodings for feature representation and is trained with mean squared error (MSE) between the reconstructed and original images in the pixel space, but only for masked patches. While MAE is also trained on RGB images only from the ImageNet-1K dataset [5], the underlying ViT architecture allows it to reason about other modalities as shown by MultiMAE [1]. Therefore, we use MAE for our study and show its effectiveness for prediction and inpainting across various modalities in top-down images useful for robotic tasks, without any finetuning.

III. Experimental Setup and Evaluation

A. Setup

Our objective is to evaluate whether the pre-trained MAE can be directly transferred from first-person RGB images to top-down RGB and semantic segmentation images as an inpainting approach. We use the MAE based on ViT-Large and perform an evaluation on the dataset collected from two

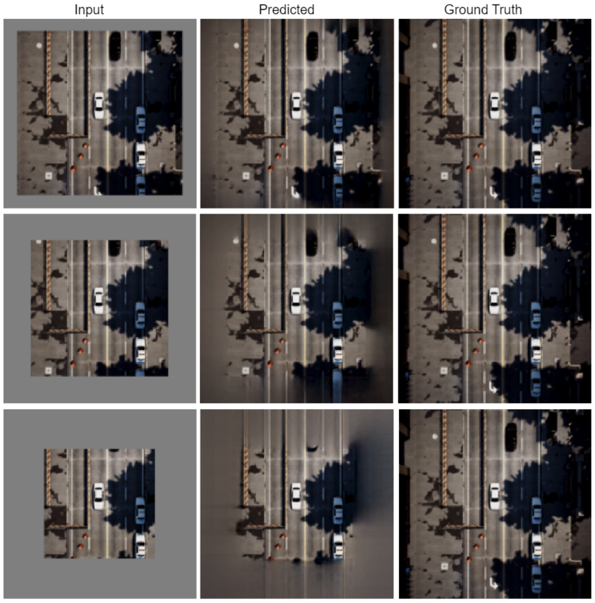

Fig. 3: Results of increasing FoV for outdoor images in three masking scenarios. MAE predicts the road well for limited masking but prediction deteriorates with larger masks.

photorealistic simulated environments, consisting of diverse indoor and outdoor scenes. The inputs to MAE are fed as RGB images by replacing labels with colors in semantic and binary maps and converting the colors in the output images back to labels by replacing them with the label assigned to the closest color in the input images.

Indoor Data: For indoor environment, we use AI2-THOR [17] which has 120 indoor scenes such as kitchens, living rooms, bathrooms, etc. We collect 1444 RGB and segmentation images with a top-own camera of a field of view of 80 degrees and rotated at intervals of 30 degrees (e.g., 30, 60, 90, etc.).

Outdoor Data: For outdoor images were taken from AirSim VALID dataset [3] which consist of scenes from cities, suburbs, and mountains among others, captured at different altitudes from an aerial robot. We sample 1000 images from this dataset for this study. For these environments, we also evaluate MAE on binary images, consisting of navigable and non-navigable regions, as a stand-in for occupancy maps.

Tasks: Inpainting in robotics tasks is distinct from typical inpainting applications as in the former we usually know what to inpaint. We study the following such applications:

1) *Increasing the FoV*: As shown by the previous works [14], [15], increasing the FoV help in long-term planning. We evaluate similarly by predicting a larger area around the image. For this, we mask the 26%, 49%, and 67% in the periphery of the image and refer to these as *Border 1*, *Border 2*, and *Border 3*, respectively.
2) *Semantics-Guided Inpainting*: The images captured during the data collection and surveying may contain undesired objects. Sometimes this may affect planning. For example, if a tree occludes the road in an aerial image, the road may be considered as *blocked* in the semantic map and hence would not be used for navigation (Figure 6). Such occlusions can be removed by MAE. While the same could be achieved with fine-tuned networks, it would be economical to use the same network that is being used for increasing the field-of-view, without task-specific fine-tuning.

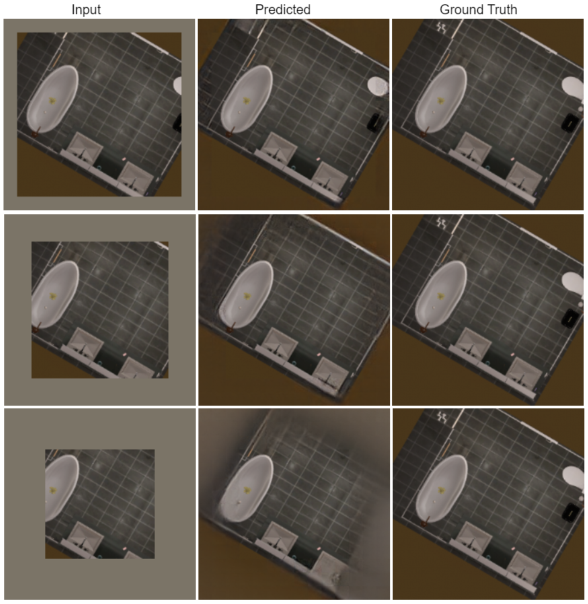

Fig. 4: Results of increasing FoV for indoor images in three masking scenarios. The corner of the bathtub and room is accurately predicted based on the symmetry of the lines.

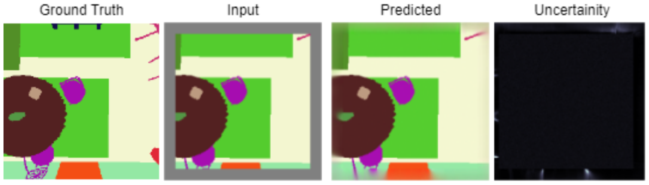

Fig. 5: Uncertainty extraction from MAE predictions.

In addition, we also present a method to extract uncertainty in predictions for uncertainty-guided navigation and exploration, as proposed in previous works [13], [23], [9].

We evaluate the RGB predictions for the FoV increase on the following metrics typically used to quantify visual similarity: (1) Frechet Inception Distance (FID), (2) Structural Similarity Index Measure (SSIM), (3) Peak Signal-to-Noise Ratio (PSNR), and (4) Mean Squared Error (MSE). For the semantic and binary images, we use Jaccard Index, i.e. mean Intersection-over-Union (mIoU) as the key metric but also provide the results for some of the aforementioned metrics since we use MAE to predict visually similar images for these modalities. For the rest of the tasks, we present qualitative examples.

B. Results

Table I summarizes the results for RGB images for both types of environments. We find that increasing the border size

(FoV) results in worse results than expected since MAE, an inpainting network can not reliably predict the outside areas without much context. Border 3 is the extreme case where the predictions get blurry. Figure 3 and Figure 4 show some examples in RGB outdoor and indoor scenes respectively and highlight this effect.

Table II and Table III summarize results for semantic and binary maps. The mIoU is very high for border 1 and goes down with increasing FoV. The effect is worse indoors as it contains many more classes (270) compared to the outdoors (30) and thus may not reliably perform color-to-label matching. Also, small objects are within the scene and on the periphery, and MAE can not predict them without seeing some part of them. Note that the Jaccard index here is not weighted by the labels' population size. Predictions on binary maps are relatively more robust since the size of objects in each class and the difference in color mapping are larger than the semantic maps. These results present an encouraging picture for a network that was not trained on such images. Figure 1 shows examples for each type of image. We share more qualitative results in the Appendix.

Uncertainty extraction can be useful for guiding navigation and exploration. MAE is a deterministic model. To extract uncertainty, we get multiple outputs by perturbing the input image with random noise. Figure 5 shows one such example where we get 10 predictions and show mean and variance in the output images. The highest variance lies in the prediction of small segments like edges and corners, and the robot should move towards these areas to get more observations. MAE is confident in its predictions about larger objects. We believe more such methods can be developed for uncertainty extraction without any change in the weight of MAE.

TABLE I: Results for increasing the FoV in RGB images

Environment	Masking	FID ↓	SSIM ↑	PSNR ↑	MSE ↓
Indoor	Border 1	**17.83**	**0.94**	**27.76**	**13.76**
	Border 2	41.79	0.86	22.23	32.42
	Border 3	76.59	0.78	19.18	52.98
Outdoor	Border 1	**53.66**	**0.84**	**26.38**	**33.59**
	Border 2	77.91	0.69	22.79	49.91
	Border 3	116.09	0.55	19.98	67.80

TABLE II: Results for increasing the FoV in Semantic segmentation images

Environment	Masking	mIoU ↑	FID ↓	SSIM ↑	PSNR ↑
Indoor	Border 1	**0.86**	**43.48**	**0.94**	**23.06**
	Border 2	0.55	75.42	0.84	17.33
	Border 3	0.34	110.01	0.78	14.90
Outdoor	Border 1	**0.90**	**42.63**	**0.94**	**25.96**
	Border 2	0.73	73.03	0.86	21.39
	Border 3	0.57	118.56	0.79	18.80

We also ran MAE on the map obtained from a TurtleBot2 robot equipped with a Hokuyo 2D scanner with an FoV of 270°. The predictions and the ground truth map for some sequences are shown in Figure 2. The robot is moving towards right here. Here MAE is used for Border-1 prediction and shows potential for use in occupancy prediction for navigation.

TABLE III: Results for increasing the FoV in Binary images from Outdoor environment

Masking	mIoU ↑	FID ↓	SSIM ↑	PSNR ↑
Border 1	**0.90**	**51.87**	**0.95**	**30.36**
Border 2	0.78	88.44	0.76	22.05
Border 3	0.64	120.94	0.56	17.81

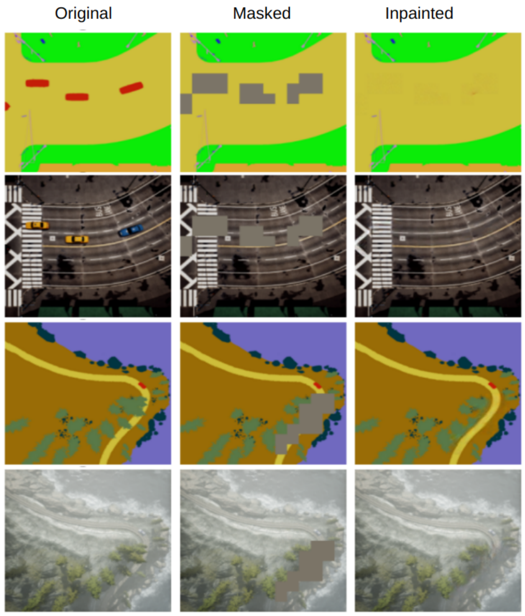

Fig. 6: Semantics guided inpainting for removing undesired objects.

Figure 6 shows some examples of semantics-guided inpainting with MAE by generating masks based on semantic maps. The first two rows remove cars from a road, to generate empty cities for digital mapping and dataset augmentations [2]. The latter two remove a tree close to the road to prescribe the path for navigation. We note that the results for the RGB image in the last example do not look visually appealing because MAE does not work well when small, non-frequent features appear in the images. However, it is noteworthy that the inpainting results for the semantic map are quite accurate.

IV. Limitations and Future Work

Although the predictions are close to the ground truth with border 1 and border 2, they significantly lack quality when more than 60% of the image is masked. This prediction can be improved if the network is fine-tuned on top-down view images. On the other hand, the network hallucinates the prediction, so if no part of an object is visible, it won't be predicted, therefore there are some cases where it fails to predict any object at all. Our future work will focus on implementing prediction-driven navigation and exploration and comparing against existing methods to quantify the efficacy of the proposed approach.

References

[1] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multimae: Multi-modal multi-task masked autoencoders. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVII*, pages 348–367. Springer, 2022.

[2] Berta Bescos, José Neira, Roland Siegwart, and Cesar Cadena. Empty Cities: Image Inpainting for a Dynamic-Object-Invariant Space. 2019.

[3] Lyujie Chen, Feng Liu, Yan Zhao, Wufan Wang, Xiaming Yuan, and Jihong Zhu. Valid: A comprehensive virtual aerial image dataset. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2009–2016, 2020.

[4] Jaime del Cerro, Christyan Cruz Ulloa, Antonio Barrientos, and Jorge de León Rivas. Unmanned aerial vehicles in agriculture: A survey. *Agronomy*, 11(2):203, 2021.

[5] 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.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[8] Omar Elharrouss, Noor Almaadeed, Somaya Al-Maadeed, and Younes Akbari. Image inpainting: A review. *Neural Processing Letters*, 51:2007–2028, 2020.

[9] Georgios Georgakis, Bernadette Bucher, Anton Arapin, Karl Schmeckpeper, Nikolai Matni, and Kostas Daniilidis. Uncertainty-driven planner for exploration and navigation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 11295–11302. IEEE, 2022.

[10] Georgios Georgakis, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, and Kostas Daniilidis. Learning to map for active semantic goal navigation. *arXiv preprint arXiv:2106.15648*, 2021.

[11] Georgios Georgakis, Karl Schmeckpeper, Karan Wanchoo, Soham Dan, Eleni Miltsakaki, Dan Roth, and Kostas Daniilidis. Cross-modal map learning for vision and language navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15460–15470, 2022.

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[13] Kapil Katyal, Katie Popek, Chris Paxton, Phil Burlina, and Gregory D Hager. Uncertainty-aware occupancy map prediction using generative networks for robot navigation. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 5453–5459. IEEE, 2019.

[14] Kapil Katyal, Katie Popek, Chris Paxton, Joseph Moore, Kevin Wolfe, Philippe Burlina, and Gregory D Hager. Occupancy map prediction using generative and fully convolutional networks for vehicle navigation. *arXiv preprint arXiv:1803.02007*, 2018.

[15] Kapil D. Katyal, Adam Polevoy, Joseph Moore, Craig Knuth, and Katie M. Popek. High-speed robot navigation using predicted occupancy maps. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5476–5482, 2021.

[16] Jing Yu Koh, Harsh Agrawal, Dhruv Batra, Richard Tucker, Austin Waters, Honglak Lee, Yinfei Yang, Jason Baldridge, and Peter Anderson. Simple and effective synthesis of indoor 3d scenes. *arXiv preprint arXiv:2204.02960*, 2022.

[17] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Matt Deitke, Kiana Ehsani, Daniel Gordon, Yuke Zhu, et al. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017.

[18] Nader Mohamed, Jameela Al-Jaroodi, Imad Jawhar, Ahmed Idries, and Farhan Mohammed. Unmanned aerial vehicles applications in future smart cities. *Technological forecasting and social change*, 153:119293, 2020.

[19] Santhosh K Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 400–418. Springer, 2020.

[20] Xuanchi Ren and Xiaolong Wang. Look outside the room: Synthesizing a consistent long-term 3d scene video from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3563–3573, 2022.

[21] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14356–14366, 2021.

[22] Vishnu Dutt Sharma, Jingxi Chen, Abhinav Shrivastava, and Pratap Tokekar. Occupancy map prediction for improved indoor robot navigation. *arXiv preprint arXiv:2203.04177*, 2022.

[23] Vishnu D Sharma, Maymoonah Toubeh, Lifeng Zhou, and Pratap Tokekar. Risk-aware planning and assignment for ground vehicles using uncertain perception from aerial vehicles. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 11763–11769. IEEE, 2020.

[24] Minghan Wei, Daewon Lee, Volkan Isler, and Daniel Lee. Occupancy map inpainting for online robot navigation. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8551–8557. IEEE, 2021.

APPENDIX

We present the results for predictions by MAE with different masking modes on RGB, semantic, and binary images in the pages below. These results are also available on our project webpage[1].

[1] Project Webpage: https://raaslab.org/projects/FLIP-TD/

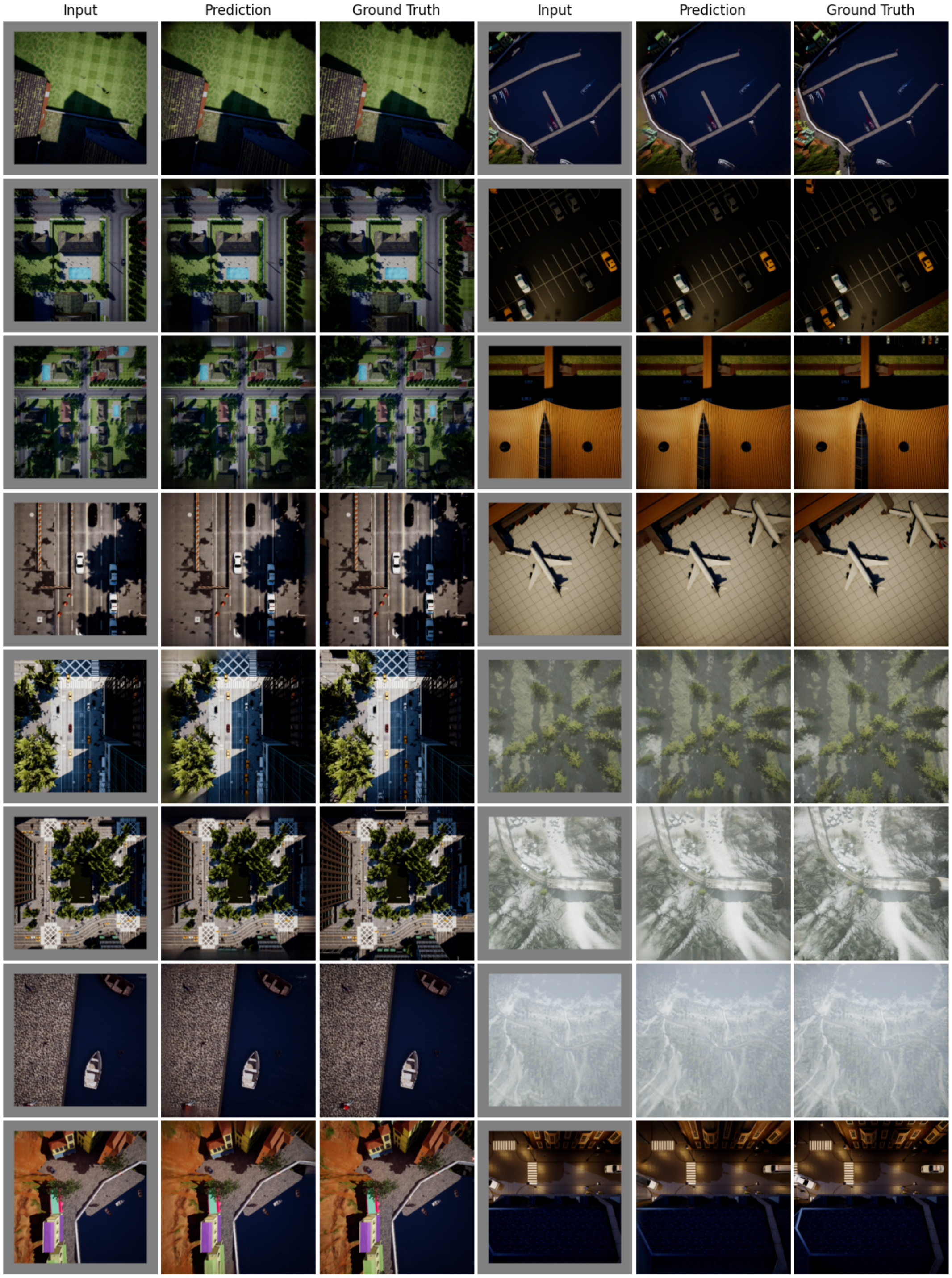

Fig. 7: Inpainting with border 1 on outdoors RGB images

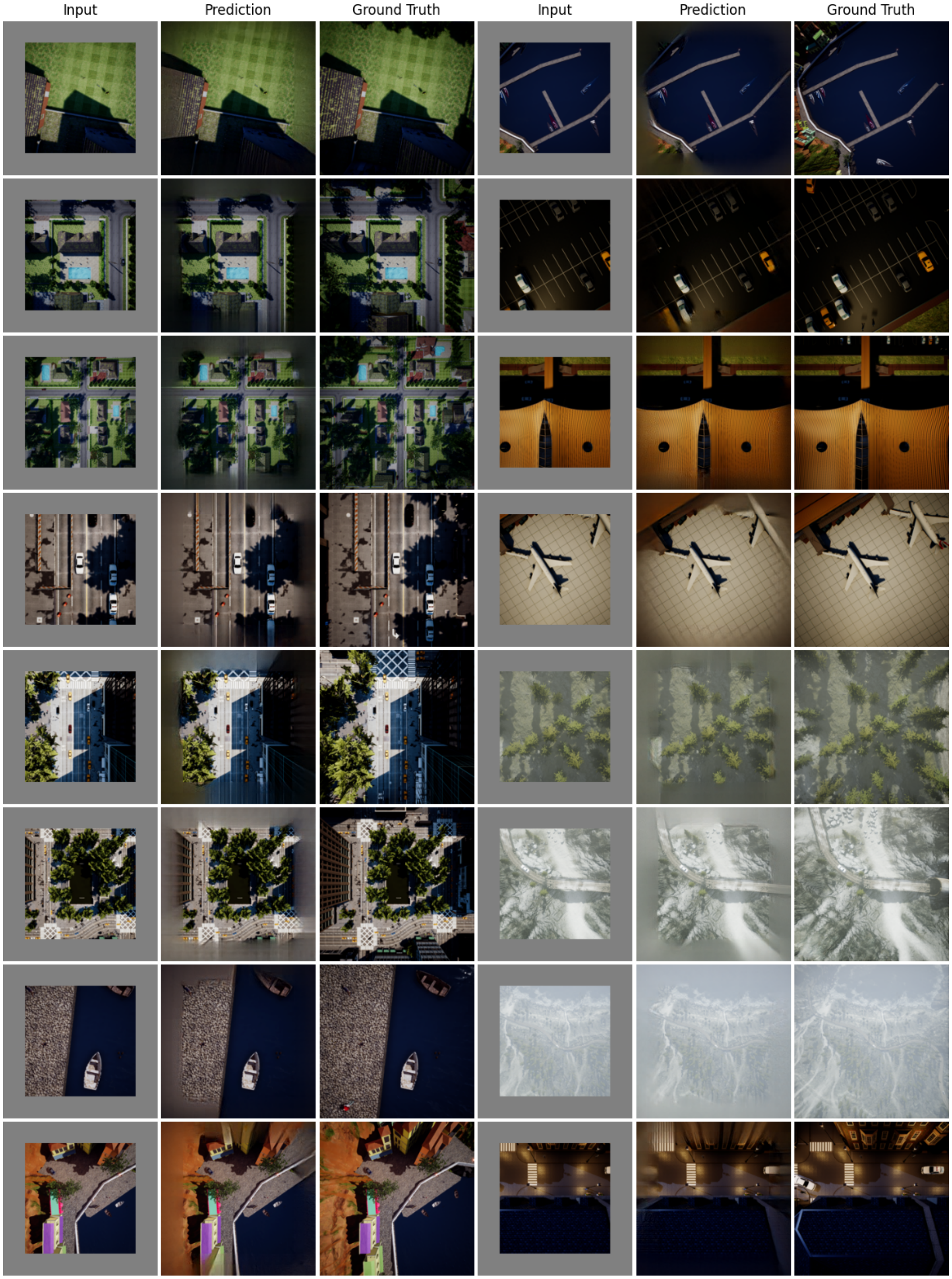

Fig. 8: Inpainting with border 2 on outdoors RGB images

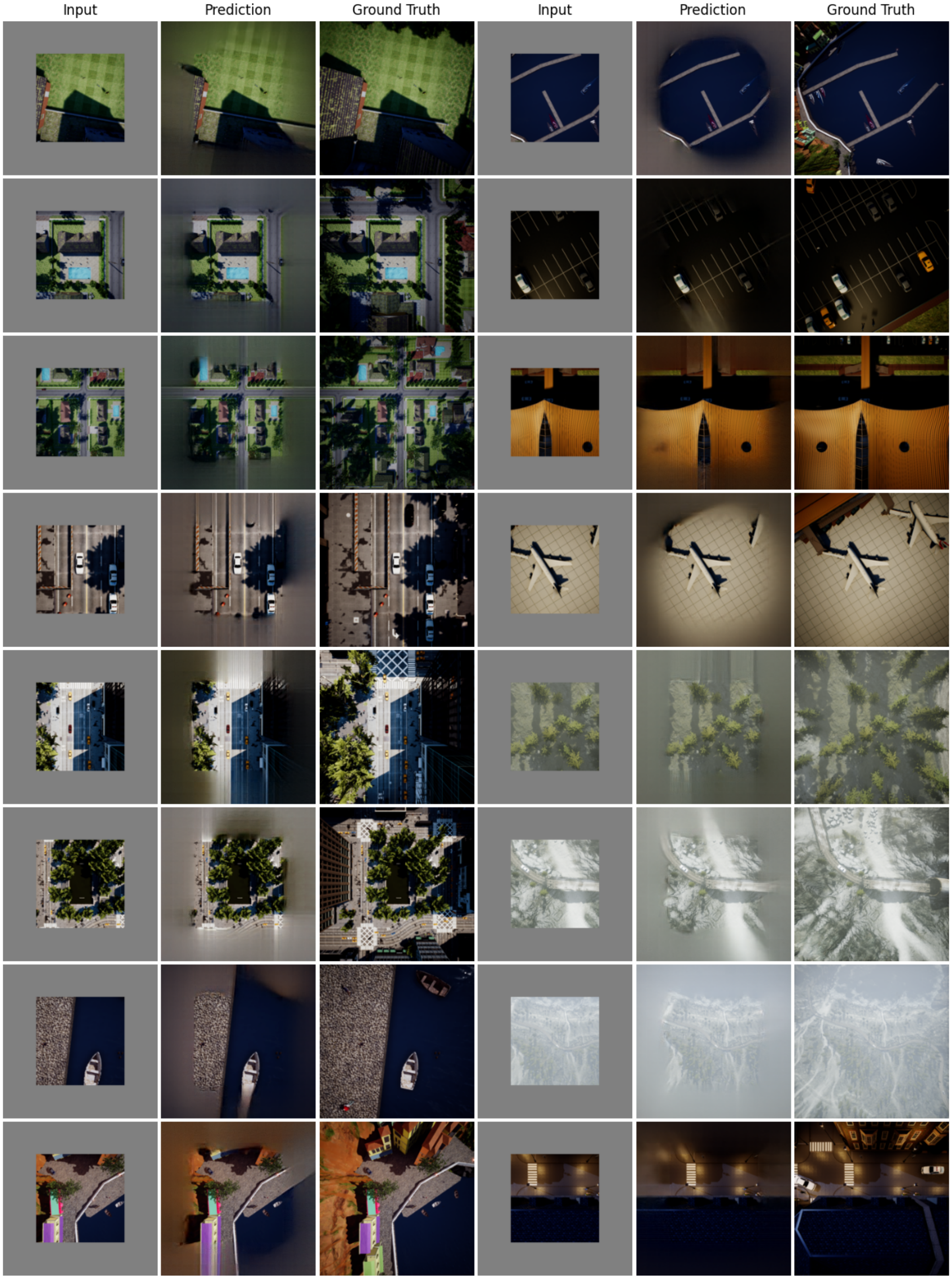

Fig. 9: Inpainting with border 3 on outdoors RGB images

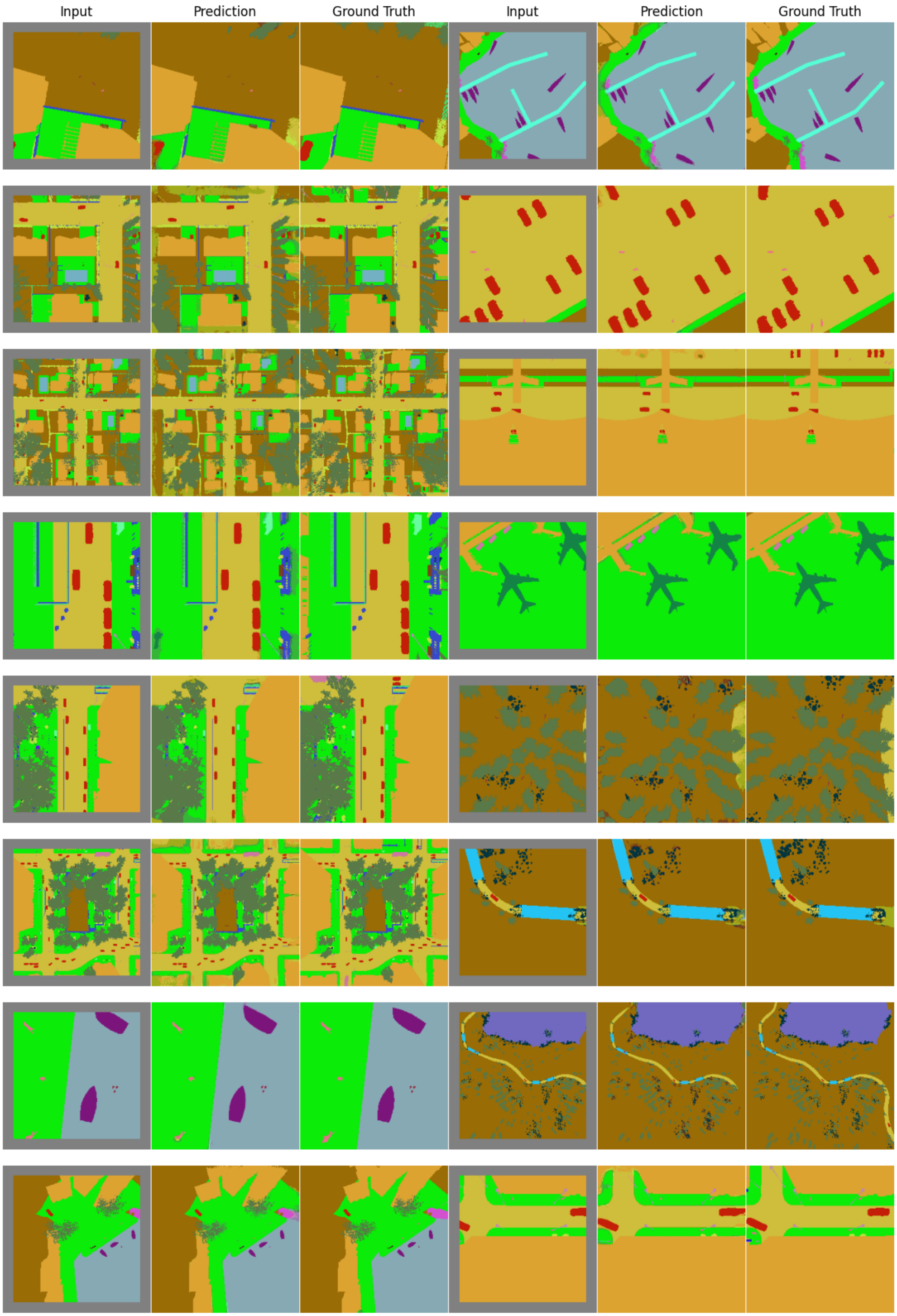

Fig. 10: Inpainting with border 1 on outdoors Semantic Segmentation Maps

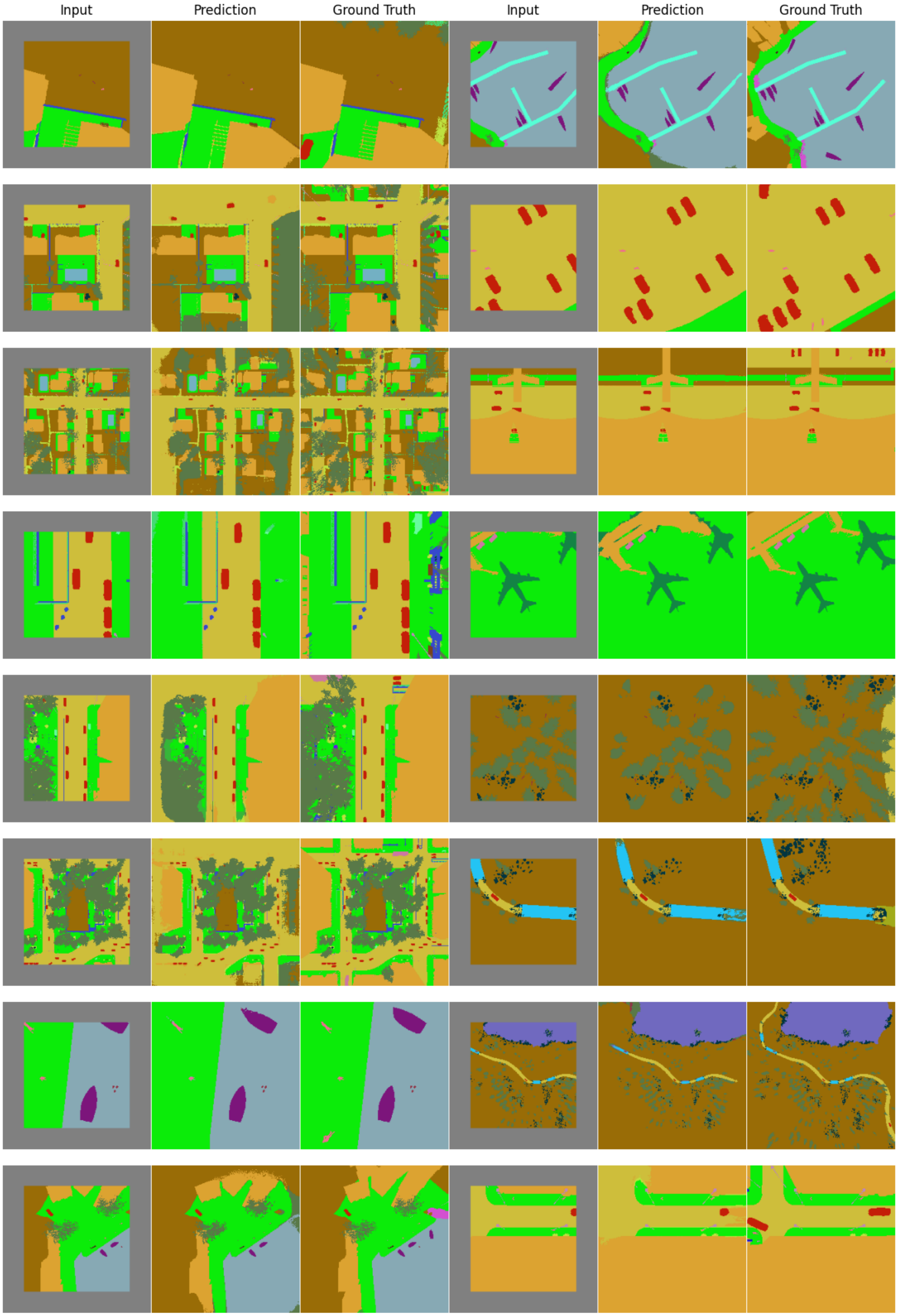

Fig. 11: Inpainting with border 2 on outdoors Semantic Segmentation Maps

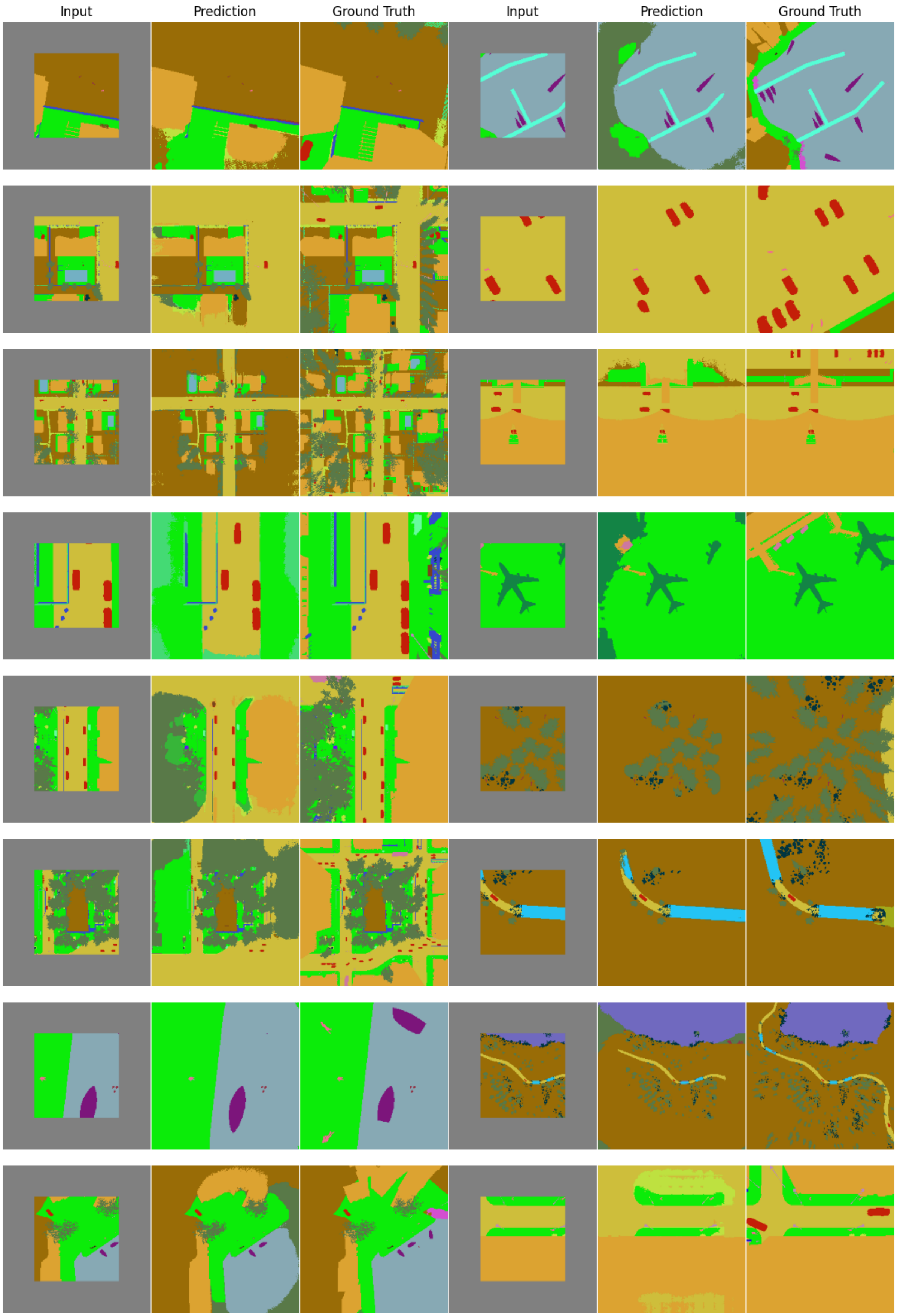

Fig. 12: Inpainting with border 3 on outdoors Semantic Segmentation Maps

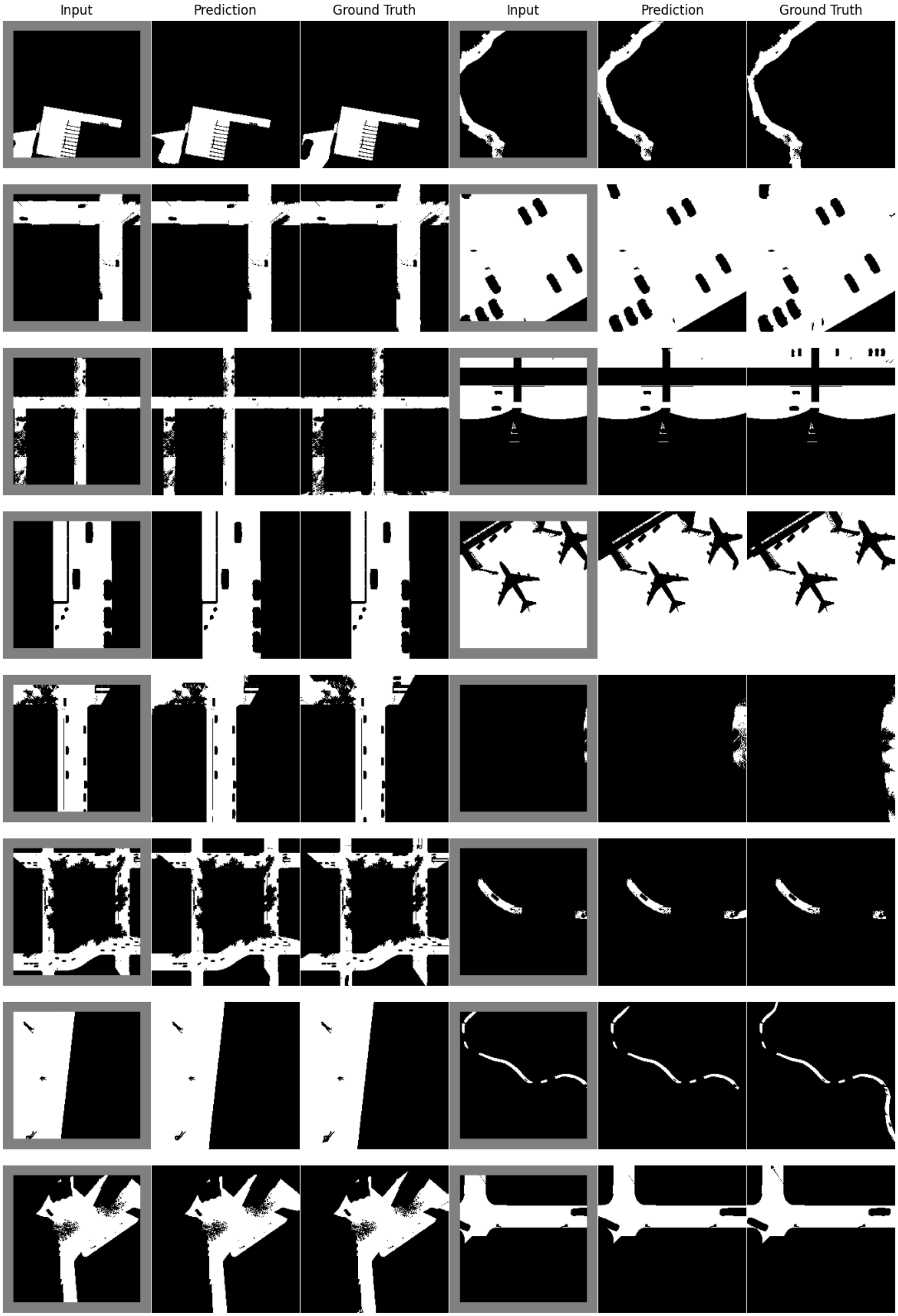

Fig. 13: Inpainting with border 1 on outdoors Binary Maps

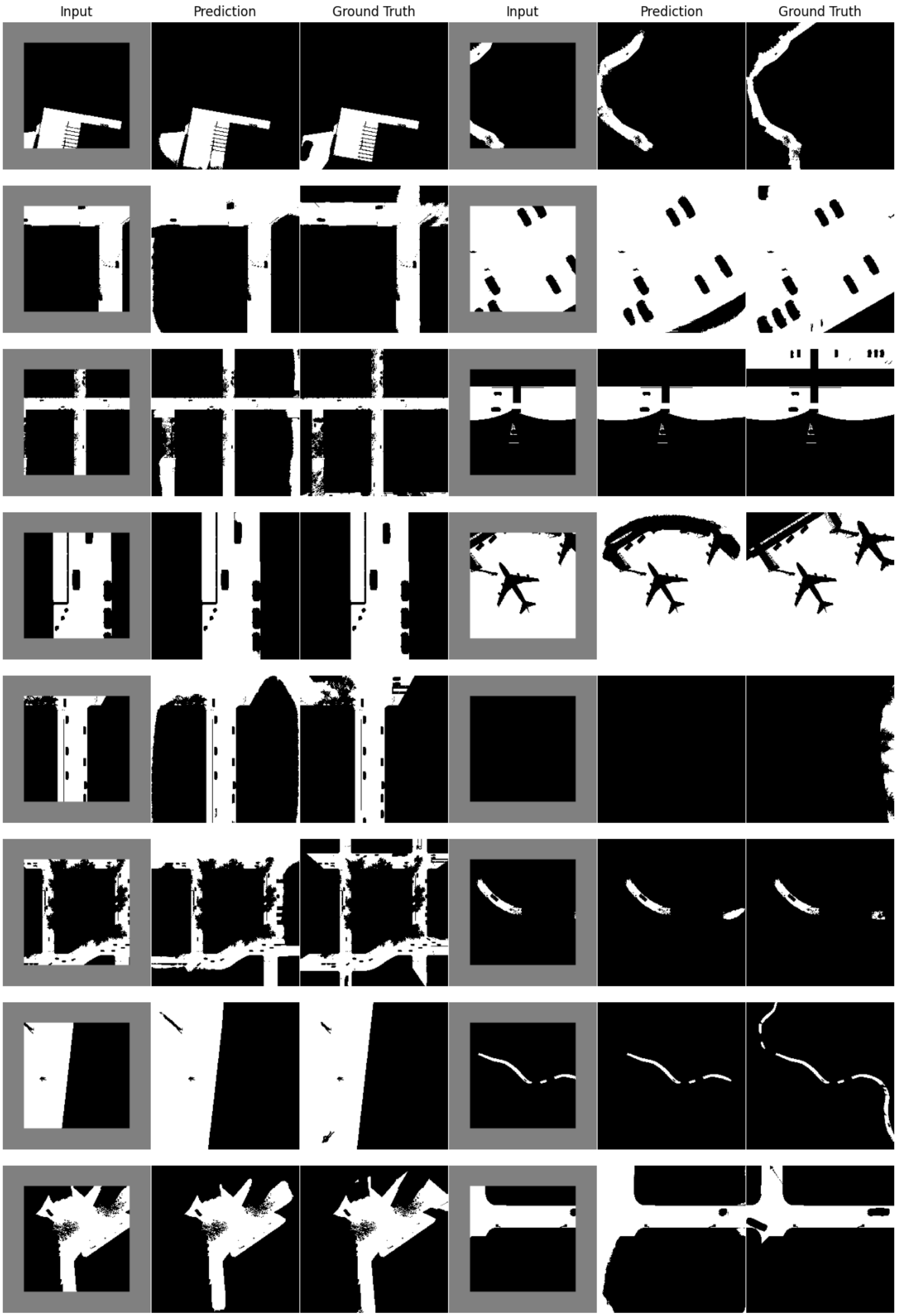

Fig. 14: Inpainting with border 2 on outdoors Binary Maps

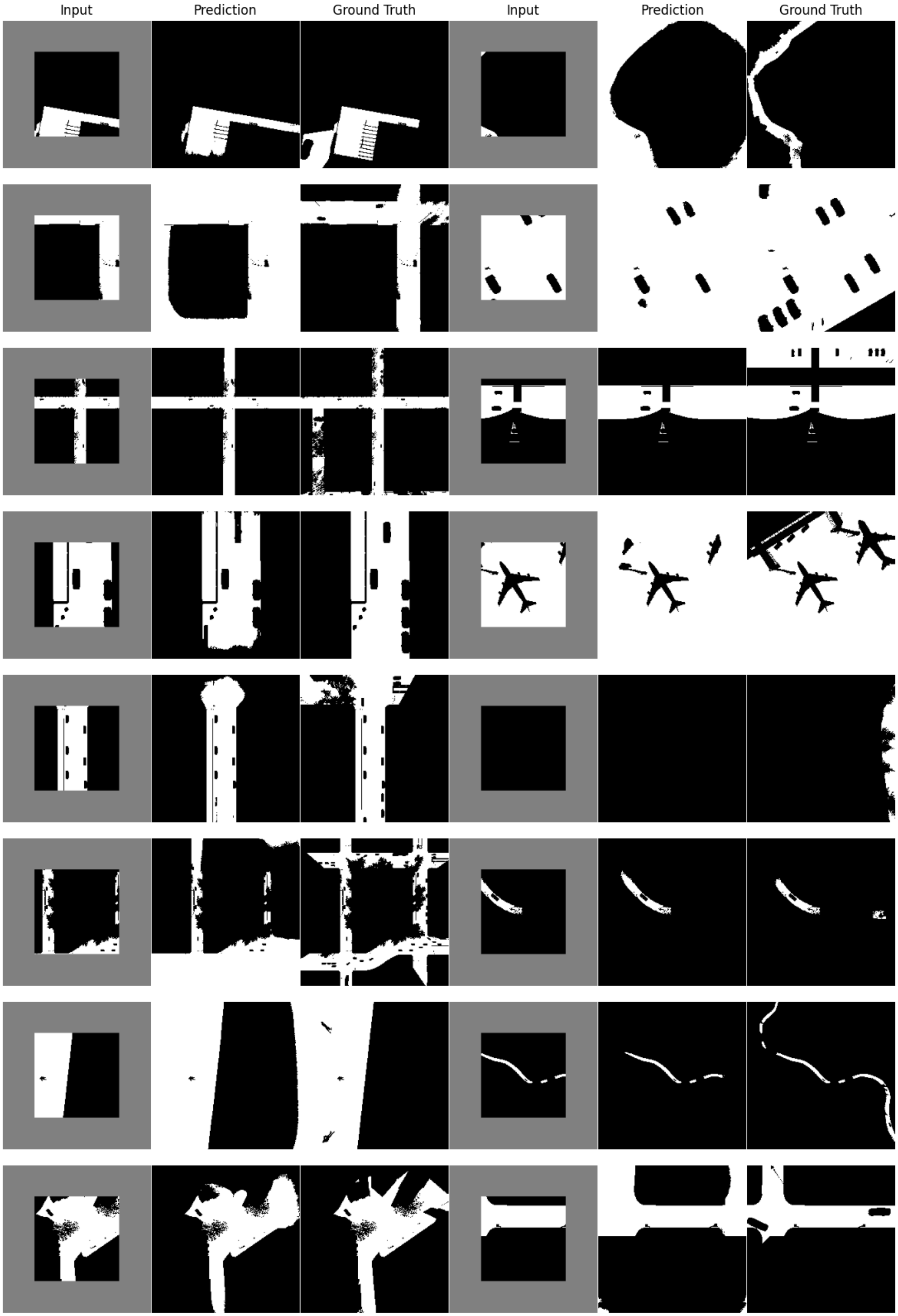

Fig. 15: Inpainting with border 3 on outdoors Binary Maps

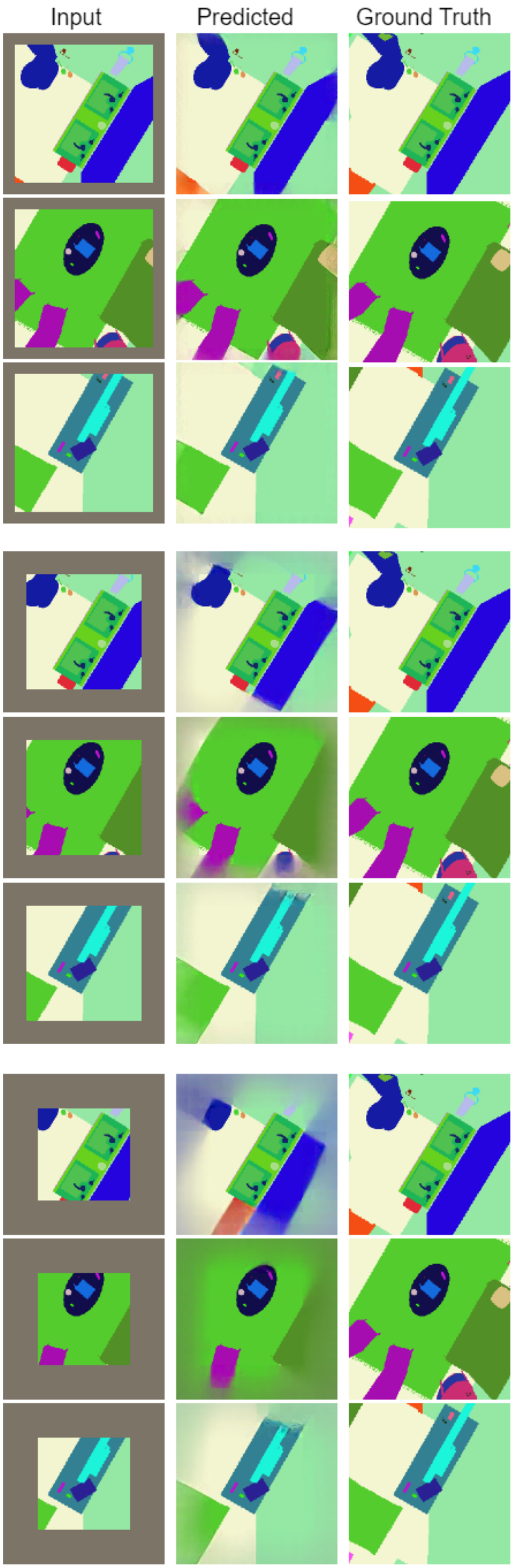

Fig. 16: Inpainting with multiple borders on indoors semantic map