VEHICLE OCCLUSION REMOVAL FROM SINGLE AERIAL IMAGES USING GENERATIVE ADVERSARIAL NETWORKS

Meijie Xiang[1], Seyedmajid Azimi[2]*, Reza Bahmanyar[2], Uwe Sörgel[1], Peter Reinartz[2]

[1]Institute for Photogrammetry, University of Stuttgart, Stuttgart, Germany
[2]Remote Sensing Technology Institute, German Aerospace Center (DLR), Oberpfaffenhofen, Germany
- xiangmeijie@foxmail.com, soergel@ifp.uni-stuttgart.de, (seyedmajid.azimi, reza.bahmanyar, peter.reinartz)@dlr.de

KEY WORDS: Aerial imagery, Deep learning, Generative Adversarial Network (GAN), HD maps, Vehicle occlusion removal.

ABSTRACT:

Removing occluding objects such as vehicles from drivable areas allows precise extraction of road boundaries and related semantic objects such as lane-markings, which is crucial for several applications such as generating high-definition maps for autonomous driving. Conventionally, multiple images of the same area taken at different times or from various perspectives are used to remove occlusions and to reconstruct the occluded areas. Nevertheless, these approaches require large amounts of data, which are not always available. Furthermore, they do not work for static occlusions caused by, among others, parked vehicles. In this paper, we address occlusion removal based on single aerial images using generative adversarial networks (GANs), which are able to deal with the mentioned challenges. To this end, we adapt several state-of-the-art GAN-based image inpainting algorithms to reconstruct the missing information. Results indicate that the StructureFlow algorithm outperforms the competitors and the restorations obtained are robust, with high visual fidelity in real-world applications. Furthermore, due to the lack of annotated aerial vehicle removal datasets, we generate a new dataset for training and validating the algorithms, the Aerial Vehicle Occlusion Removal (AVOR) dataset. To the best of our knowledge, our work is the first to address vehicle removal using deep learning algorithms to enhance maps.

1. INTRODUCTION

With the rapid evolution of autonomous driving, there has been a rising demand for high-definition (HD) maps in recent years. Occlusion-free aerial images from drivable areas can help generating more precise and complete HD maps by allowing for more accurate extraction of crucial features such as road boundaries and lane markings. Automatically removing occlusions is carried out by first detecting and masking undesired occlusions caused by static and dynamic objects such as vehicles, and then reconstructing the missing information in the masked areas, with both of these tasks being non-trivial. Several previous works using classical and learning-based approaches have addressed occlusion removal as an inpainting problem, filling in missing areas with the support of known surrounding areas. While classical methods rely solely on the neighborhoods of the missing areas, learning-based approaches are capable of using the learned features from various similar images. This theoretically allows them to restore features that are unrelated to the neighbouring regions.

Among the learning-based approaches, the ones based on deep learning (DL) has shown promising performance in various image processing and computer vision tasks in the past two decades. For the first time, (Pathak et al., 2016) employed a DL-based approach for image inpainting. They proposed a generative adversarial network (GAN), where the encoder-decoder structure can generate the incomplete image parts.

Later, (Nazeri et al., 2019) proposed the Edge-Connect network, focusing on restoring the missing image structural features by two GANs: one for generating edges, and the other one for completing the missing areas. The first GAN takes an original image and its corresponding missing information mask as input, and generates a masked grayscale image and a masked edge map. For the training step, the ground truth edge map is generated by applying the Canny edge detector to the original image. The resulting edges are then transferred to the second GAN with the masked source image in order to obtain the final output. In another work, (Ren et al., 2019) proposed the StructureFlow network which uses Edge-Connect as backbone network for recovering the missing structures in two stages: generating edge-preserved smooth images and refining the uniformity of textures. The inputs to the network include the source image and its corresponding mask, as well as its masked structure map. In contrast to Edge-Connect, StructureFlow attempts to recover a smoothed structure map rather than an edge map for structure retention, resulting in a better reconstruction of structures and textures.

As a more efficient method, (Li et al., 2019) proposed progressive reconstruction of visual structure (PRVS), which performs structure and texture restorations in parallel. The encoder starts the process from the boundaries of the masked area to its center, while the decoder performs the same operation in the opposite direction. This procedure enhances the coherence between the masked area and the rest of the image, considering that the boundaries of the masked area provide valuable information. Results on several benchmark datasets show promising restorations of image contents and edges. Mutual encoder-decoder with feature equalization (MEDFE) is another method that simultaneously recovers structural and textural information (Hongyu Liu, Yang, 2020). This network recovers textural and structural features at the shallowest and deepest layers, respectively, during the encoding process, and adds them through skip connections during the decoding step.

*Corresponding author

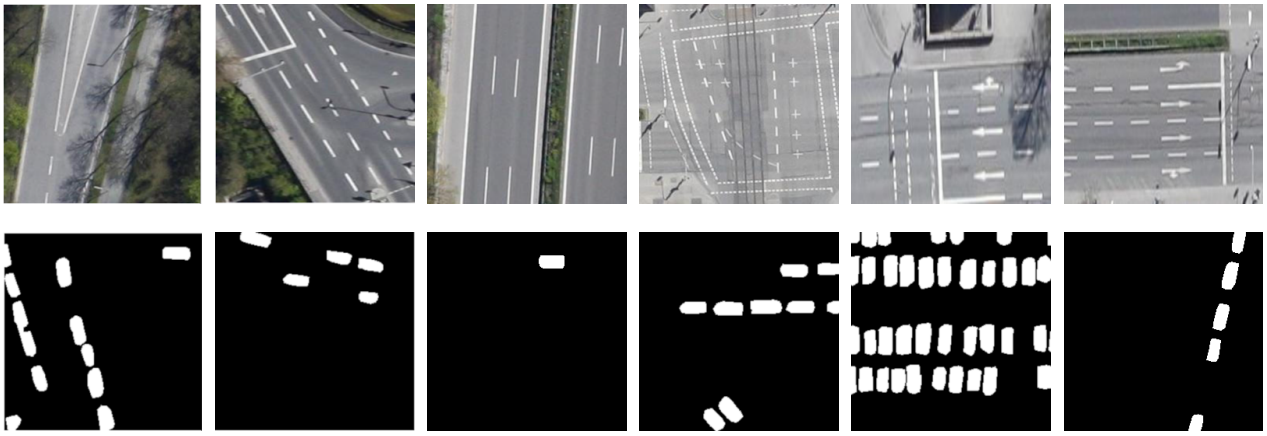


Figure 1. Examples of occlusion-free images and random occlusion masks from the AVOR dataset.

In order to improve pixel discontinuity in the missing regions, (Liu et al., 2019) developed the coherent semantic attention (CSA) network, consisting of a shallow and a deep network. The shallow network provides a course prediction of the restored image, while the deep network refines the output using the CSA layer to improve the pixel coherence in the missing areas. Results show that CSA can coherently recover missing regions. Dealing with restoring large missing regions, (Li et al., 2020) proposed the recurrent feature reasoning (RFR) network that recursively infers the boundary of the missing regions as a reference, filling the missing parts from the border towards the center. RFR uses a knowledge consistent attention module that calculates the scores for each recursion and merges them in order to obtain more consistent outputs. Results demonstrate its efficiency in reconstructing large missing regions.

In real-world scenarios, there is no ground truth for missing image information, with inpainting methods estimating the possible restorations of the missing information. Therefore, there is no unique solution to inpainting problems. In order to deal with these limitations, a more recent family of inpainting methods tries to provide multiple possible restorations of the missing image parts. One of such methods is the hierarchical vector quantized variational auto-encoder (VQ-VAE) network (Peng et al., 2021) that comprises three modules: a hierarchical encoder and decoder extracting discrete structural and textual features, a diverse structure generator estimating structure distribution in order to produce multiple possible structural features, and a texture generator which helps maintaining the synthesis of textures.

All these methods have been developed for image inpainting and occlusion removal in the computer vision frame. To the best of our knowledge, there is no DL-based method in remote sensing for removing occlusion from single high-resolution aerial images. Moreover, while datasets with ground truth are crucial for training DL-based methods, there is no training dataset available for removing vehicles from aerial images. Since usually there is no ground truth of the occluded image parts, generating such dataset is very challenging thing, which imposes limits to the development of DL-based methods in this domain.

In order to deal with these limitations and to promote the future development of DL-based methods for vehicle removal from single aerial images, in this paper we introduce the Aerial Vehicle Occlusion Removal (AVOR) dataset, based on an aerial image dataset with annotated vehicles from the German Aerospace Center (DLR), the so-called DLR multi-class vehicle detection and orientation in aerial imagery (DLR-MVDA) (Liu, Mattyus, 2015). We consider only vehicles as occluding objects, without their corresponding shadows. Our Occlusion-free dataset is composed of 1,296 images of size 256 × 256 pixels, containing no vehicle occlusions. Furthermore, in order to train the DL networks to learn the occlusions, we generate 19,639 realistic vehicle occlusion masks with the same size as the images. We randomly assign the masks to the images, then we split the dataset into training, validation, and test sets. The number of masks and images are equal to the number of images with a fixed assignment in the test set. However, for the training and validation sets, the number of masks are larger than those of the images, and we perform an on-the-fly assignment during the training and validation phases. Figure 1 demonstrates some example images and occlusion masks from the AVOR dataset. Furthermore, as an additional contribution, we adapt the aforementioned state-of-the-art GAN-based inpainting methods and apply them on our AVOR dataset. As demonstrated in Figure 2, the training and inference procedures of GAN-based techniques exhibit similarities, despite variations in their structural specifics. We then investigate their performances qualitatively and quantitatively, and discuss their opportunities and limitations for their practical use in future applications by the community. According to the presented results, StructureFlow outperforms the other methods and its restorations are robust with high visual fidelity in real-world applications.

2. AERIAL VEHICLE OCCLUSION REMOVAL DATASET

In this section, we introduce our Aerial Vehicle Occlusion Removal (AVOR) dataset. We generated AVOR based on DLR-MVDA, an aerial image dataset with annotated vehicles (Liu, Mattyus, 2015), which comprises 20 high-resolution and non-overlapping aerial RGB images with a size of 5616×3744 pixels taken during a flight campaign over Munich, Germany by a helicopter. The images were acquired at 1000 m resulting in a ground sampling distance (GSD) of 13 cm/pixel.

Similar to many annotated datasets, this dataset suffers from a lack of ground truth for the occluded regions, which is crucial for training purposes. The general idea is to extract occlusion-free areas from original images for driving areas which are not occluded by vehicles. In order to train the algorithms to learn

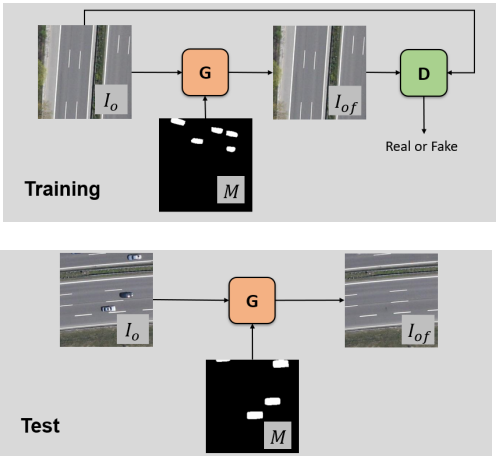


Figure 2. Workflow of GAN-based inpainting algorithms including the training and test phases.

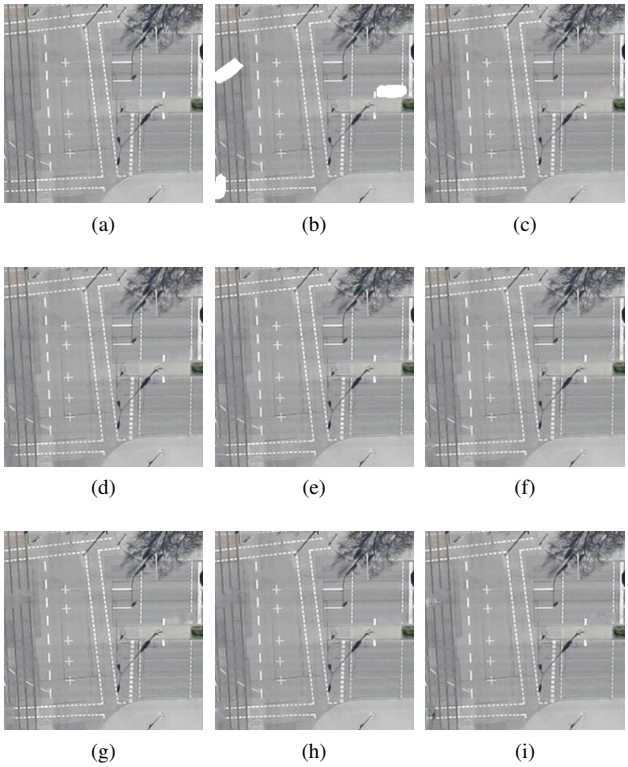


Figure 3. Examples of image restoration on the AVOR dataset. The original occlusion-free (a) and masked images (b). The results by Edge-Connect (c), StructureFlow (d), MEDFE (e), PRVS (f), CSA (g), RFR (h), and VQ-VAE (i).

the occluded image contents, we generate synthetic occlusion masks and assign them to these occlusion-free images. Since the input is the element-wise multiplication of source images and their binary masks, the networks assume that the masked parts of the original images are occluded by vehicles.

In order to generate the occlusion-free images, we manually crop image subsets of 256×256 pixels containing drivable areas not occluded by vehicles from the large aerial images. The patch size is a trade-off between the number of samples and the occurrence of representative contextual features. Since the number of such image patches is limited, we augment them by rotations of 90, 180, and 270 degrees, resulting in 1,296 image patches. Then we split the dataset into 1,050 training, 188 test, and 58 validation images. Figure 1 shows a few example images and occlusion masks from the AVOR dataset.

Since the images are not occluded, their corresponding occlusion masks can be randomized. In order to keep the masks as realistic as possible, we randomly crop patches of 256×256 pixels from the binary masks of the occluding vehicles of the original dataset, resulting in 19,639 masks. We then split the generated masks into 17,825 training, 188 test, and 1,626 validation masks, where the image patches of each set do not belong to the same images. The number of masks in the training and validation sets is much larger than in image patches. Thus, one occlusion-free image can match multiple masks during training, which can slightly compensate for the limited number of occlusion-free images, as various occlusion scenarios for each image are present. For the test time, we use a fixed set of 188 randomly selected masks.

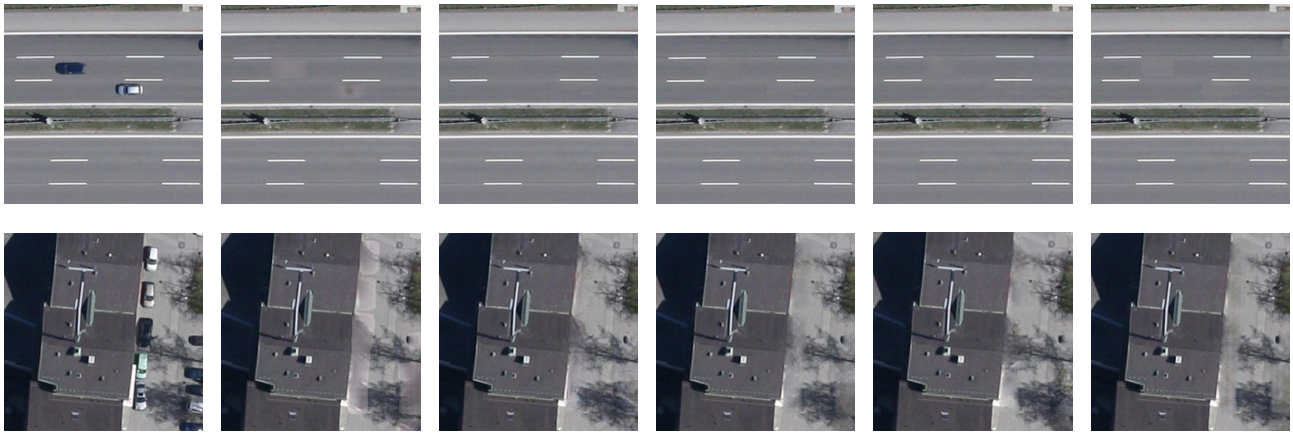
The most notable advantage of the dataset is that it contains real-world images of drivable areas without occlusion, and occlusion masks from real occluding vehicles. The dataset has also some limitations, such as its relatively small size and unbalanced distribution of scenes. For example, since the parking areas are usually occupied by vehicles, these are rarely presented in our dataset: this can have a negative impact on occlusion removal performance in these regions.

3. VEHICLE REMOVAL USING INPAINTING METHODS

In this section we report our experiments on vehicle removal using image inpainting methods and evaluate their results. The most significant challenge in vehicle removal is reconstructing the missing information. The GAN-based inpainting methods learn the data models by training on relevant datasets and use the learned model to generate the missing features. Despite the difference in their structure details, the training and inference procedures of the GAN-based methods are similar as shown in Figure 2.

In the training step, the networks relies on occlusion-free images and occlusion masks to learn the characteristics of the occluded areas by their generators. For inference, the generators reconstruct the missing information of the occluded areas indicated by the occlusion masks. In Figure 2, G and D denote generator and discriminator, respectively. I_o is the input occlusion-free image, M is a binary occlusion mask, and I_{of} is the generated occlusion-free image. The discriminator uses the original occlusion-free image (I_o) as ground truth.

For our experiments, we consider seven GAN-based methods including StructureFlow (Ren et al., 2019), Edge-Connect (Nazeri et al., 2019), PRVS (Li et al., 2019), MEDFE (Hongyu Liu, Yang, 2020), RFR (Li et al., 2020), CSA (Liu et al., 2019), and VQ-VAE (Peng et al., 2021). We use the available implementations of the algorithms available on Github, and keep their parameters and configurations as in the original networks. We train the networks on the AVOR dataset for 300 epochs. For training, we input an occlusion-free image with a randomly selected occlusion mask from the training set to the network. We evaluate the methods on the test set of the AVOR dataset both qualitatively and quantitatively. To this end, we mask the input



(a) (b) (c) (d) (e) (f)

Figure 4. Vehicle removal examples including the original images (a), and the results by Edge-Connect (b), StructureFlow (c), MEDFE (d), PRVS (e), and RFR (f).

Figure 5. Demonstration of vehicle removal by StructureFlow.

occlusion-free images by their corresponding masks, and compare the restored images with the input occlusion-free images (ground truth). Furthermore, we apply the trained models to the real-world scenarios by inputting images with vehicle occlusions (from DLR-MVDA) together with the binary mask of the vehicles. The networks are supposed to replace the vehicles with the image content that they occlude. Since there is no ground truth available for the occluded images, we evaluate the results qualitatively.

4. EVALUATION METRICS

To evaluate the image restoration performance from various perspectives, we employ three commonly-used metrics in image enhancement domain.

Peak Signal-to-Noise Ratio (PSNR) (Davda et al., 2010) is the most commonly-used metric for image quality. It characterizes the relationship between the maximum possible signal power and the destructive noise power. Since signals can have wide dynamic ranges, PSNR is often presented as a logarithm with a decibel range of 0 to ∞:

$$PSNR = 10.\log\left(\frac{(\max(I))^2}{MSE}\right), \quad (1)$$

where $\max(I)$ is the maximum pixel value of the image and MSE is the mean squared error. The larger the PSNR value, the better the quality of the reconstructed image, as less errors are introduced to the output.

Structural Similarity (SSIM) (Davda et al., 2010) is derived from three comparative measures (brightness, contrast, and structure) between the source image x and the reconstructed image y as follows:

$$SSIM(x,y) = \left[l(x,y)^{\alpha} \cdot c(x,y)^{\beta} \cdot s(x,y)^{\gamma}\right], \quad (2)$$

where α, β, and γ are the weights of brightness, contrast and structure, respectively. The SSIM value ranges between 0 and 1. It is equal to 1 only if the two images are identical.

Fréchet Inception Distance (FID) (Heusel et al., 2017) is a widely-used metric for measuring the distances between the feature vectors of the original image set x and the recovered image set y, as:

$$FID(y,x) = \| \mu_y - \mu_x \|_2^2 + \mathrm{T_r}\left(\Sigma_y + \Sigma_x - 2\left(\Sigma_y\Sigma_x\right)^{\frac{1}{2}}\right), \quad (3)$$

where μ_y and μ_x denote the mean values of the feature vectors of the sets y and x, respectively. Additionally, Σ_y and Σ_x correspond to the covariance matrices of the feature vectors of the two sets, respectively, while $T_r(.)$ is the trace of the corresponding matrix. A smaller FID implies a higher similarity of the generated image to the source, with the FID between two identical images being 0.

5. RESULTS AND DISCUSSION

Table 1 shows the quantitative evaluation of the results on the test set of the AVOR dataset based on SSIM, PSNR, and FID

Algorithm	SSIM ↑	PSNR ↑	FID ↓
Edge-Connect (Nazeri et al., 2019)	0.983 (0.027)	38.71 (6.77)	15.53
StructureFlow (Ren et al., 2019)	**0.990 (0.014)**	**41.02 (6.51)**	**7.88**
MEDFE (Hongyu Liu, Yang, 2020)	0.989 (0.017)	40.48 (6.99)	9.87
PRVS (Li et al., 2019)	0.988 (0.016)	40.21 (6.68)	21.91
CSA (Liu et al., 2019)	0.985 (0.021)	38.57 (6.18)	12.47
RFR (Li et al., 2020)	0.988 (0.016)	40.29 (7.22)	34.52
VQ-VAE (Peng et al., 2021)	0.984 (0.022)	38.37 (6.91)	13.30

Table 1. Comparing occlusion removal by inpainting algorithms on the occlusion-free dataset

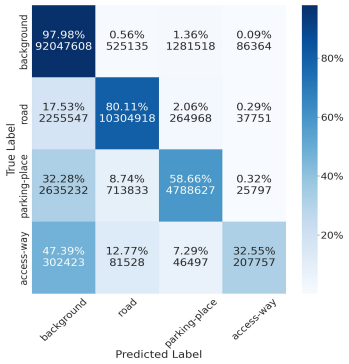
metrics. For SSIM and PSNR, we also present standard deviations in parenthesis. In this table, ↑ and ↓ mean that higher and lower indicate better performance, respectively, and we report in bold the best result for each metric. According to the results, StructureFlow (Ren et al., 2019) outperforms the other methods. It yields the best results for SSIM and PSNR, showing that it can better preserve the image structures and the overall pixel values. Moreover, it achieves the best FID, indicating that the pixel value distributions of its resulting images are close to those of the source images. In order to provide a visual indication of the capabilities of each method, we assess the results qualitatively. Figure 3 shows an example reconstruction of missing features by different methods. In this figure, the first image is an occlusion-free image selected from the test set of the AVOR dataset. The second image is the masked image which is given to the networks, and the remainder the reconstruction results.

Moreover, among the multiple generated outputs by VQ-VAE (Peng et al., 2021), we select and visualize a representative one in Figure 3. According to the results StructureFlow ensures the continuity of the missing structures, although part of its created structures are not similar to the original image. Additionally, all methods fail in properly reconstructing fine structures such as dashed lines. Only StructureFlow could partially complete a missing dashed line on the right side of the image as a continuous line.

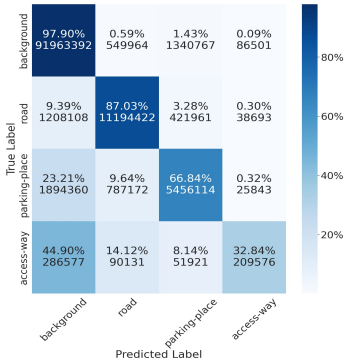
In order to evaluate the applicability of the methods on real-world vehicle removal problems, we apply the top five methods (including Edge-Connect, StructureFlow, MEDFE, PRVS, and RFR) based on the qualitative and quantitative results (see Table 1) to the original images of the DLR-MVDA dataset.

Figure 4 demonstrates the results on two example image patches with diverse occlusion scenarios. For the first example, since the occlusions lay on a homogeneous road surface and do not require much structural reconstructions, all methods can restore the missing parts in a satisfactory way. However, results obtained with Edge-Connect suffer from texture inconsistencies, indicating its limitations in preserving texture homogeneity. In the second example, the same performance holds for vehicles on the plane road surface. For vehicles occluding the tree shadow textures, StructureFlow and PRVS outperform the other methods in reconstructing the missing tree features.

In order to provide a broader view on the vehicle removal in real-world scenarios, Figure 5 represents results of StructureFlow on a large part of an image from the DLR-MVDA dataset, where StructureFlow can remove most of the vehicles and restore the occluded road information. There are also a few failure cases, especially where the vehicle shadows are large. Since the vehicle masks usually do not include shadows, the models cannot learn how to deal with the significant contrasts on the border of the missing areas which do not belong to the road



(a) Original images with vehicles



(b) After vehicle removal

Figure 6. Confusion matrices before and after the vehicle removal.

Table 2. Surface extraction evaluation for HD mapping using aerial images after vehicle removal. Numbers are in %.

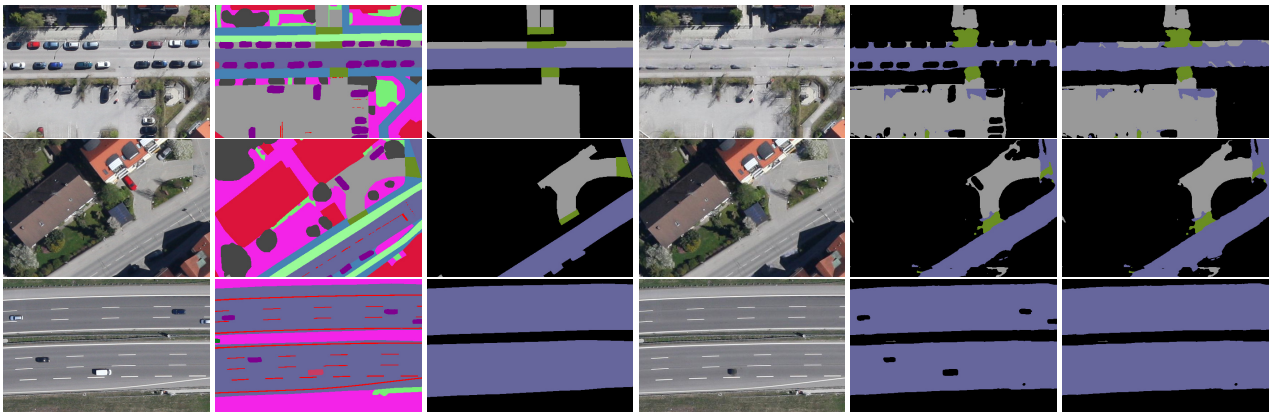
Scenario	Acc.	Precision	Recall	DICE	IoU
Vehicle	92.86	79.11	67.33	72.01	60.24
No vehicle	94.13	79.57	71.15	74.42	63.51

Table 3. Surface extraction evaluation for HD mapping using aerial images after vehicle removal. MN, BKG, RD, PP and AW/EE stand for mean, background, road, parking-place, Access-way or Entrance-exit.

Scenario	IoU %				
	MN	BKG	RD	PP	AW/EE
Vehicle	60.24	92.85	72.65	49.08	26.36
No vehicle	63.51	94.49	78.33	54.68	26.55

surface. Thus, the missing parts with the pixel values of the shadowed areas appear smeared. This shows the limitations caused by the vehicle masks, and the necessity for the development of algorithms learning the vehicles and their relevant features (e.g., shadows) in the training phase, in order to recognize and remove them fully automatically without relying on vehicle masks as prior information.

To investigate the improvements of surface extraction necessary for HD mapping in autonomous driving, we use the labels of the SkyScapes (Azimi et al., 2019) dataset. We keep only road, parking-place and enterance-exit (access-way) classes. To remove vehicles, we propagate the class of neighboring pixels based on 8-pixel connectivity to the regions occupied by vehi-



(a) Original image (b) Base GT (c) Subset GT (d) Output vehicle removal (e) Prediction with vehicles (f) Prediction without vehicles

Figure 7. Results of surface extraction for HD mapping after vehicle removal. (b) is the full set of labels from the SkyScapes dataset which refer to (Azimi et al., 2019) for its color codes. Surface color coding: road, parking-place, entrance-exit (accesss-way).

cles. In this propagation, we only allow the pixels belonging to the mentioned classes to be propagated. One could have used morphological or contour-based propagation, but we found this approach to be more accurate. We compare the predictions of SkyScapesNet (Azimi et al., 2019) with and without vehicles against the generated ground truth. Table 2 and Table 3 show how the performance can be increased on the indicated classes without vehicles. The results show after the vehicle removal, the mean IoU has increased from 60.24% to 63.51%, a roughly 3% increase, indicating the rough portion of vehicles occupying the driving areas in the predictions. The confusion matrices in Figure 6 provide more insights into the segmentation results where Figure 7 illustrates the qualitative evaluation results on three sample patches. We expect that by applying the segmentation method to the images with vehicles removed, we can achieve even better performance than what our preliminary experiment indicates.

6. CONCLUSION AND FUTURE WORK

In this paper, we address automatic vehicle removal from drivable areas in aerial imagery using DL-based inpainting methods. Due to the lack of appropriate training datasets, we generate the Aerial Vehicle Occlusion Removal (AVOR) dataset containing occlusion-free aerial image patches and realistic random vehicle occlusion masks. Subsequently, we adapt and evaluate seven state-of-the-art GAN-based inpainting methods on the AVOR dataset. Based on quantitative and qualitative evaluations, StructureFlow outperforms other inpainting methods, yielding robust restorations with high visual fidelity. Results show that all evaluated methods suffer from limitations in restoring fine structures such as lane markings on the road surfaces. In order to improve their performance, in addition to developing problem-specific networks, future work should focus on generating larger and more diverse datasets, in terms of occluding vehicles, road surface textures, and structures. Moreover, future works should make the vehicle removal algorithms independent from vehicle masks which in this stage is needed as an extra input to the networks. Using the SkyScapes dataset, we demonstrate that vehicle removal can improve the performance of surface extraction. As a next step, we plan to explore the direct application of vehicle removal to trained networks using images generated by this approach.

REFERENCES

Azimi, S. M., Henry, C., Sommer, L., Schumann, A., Vig, E., 2019. Skyscapes - fine-grained semantic understanding of aerial scenes. *Proceedings of the IEEE international conference on computer vision*.

Davda, S., Vora, V., Suthar, A., Makwana, Y., Davda, S., 2010. Analysis of Compressed Image Quality Assessments. *International Journal of Advanced Engineering Application*, 10.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Klambauer, G., Hochreiter, S., 2017. Gans trained by a two time-scale update rule converge to a nash equilibrium. *Proc. NIPS*.

Hongyu Liu, Bin Jiang, Y. S. W. H., Yang, C., 2020. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. *Proc. ECCV*.

Li, J., He, F., Zhang, L., Du, B., Tao, D., 2019. Progressive reconstruction of visual structure for image inpainting. *Proc. ICCV*.

Li, J., Wang, N., Zhang, L., Du, B., Tao, D., 2020. Recurrent feature reasoning for image inpainting. *Proc. CVPR*.

Liu, H., Jiang, B., Xiao, Y., Yang, C., 2019. Coherent semantic attention for image inpainting. *Proc. ICCV*.

Liu, K., Mattyus, G., 2015. Fast Multiclass Vehicle Detection on Aerial Images. *IEEE Geoscience and Remote Sensing Letters*, 12(9), 1938-1942.

Nazeri, K., Ng, E., Joseph, T., Qureshi, F. Z., Ebrahimi, M., 2019. EdgeConnect: Generative Image Inpainting with Adversarial Edge Learning. *arXiv: 1901.00212*.

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A. A., 2016. Context Encoders: Feature Learning by Inpainting. *arXiv: 1604.07379*.

Peng, J., Liu, D., Xu, S., Li, H., 2021. Generating diverse structure for image inpainting with hierarchical vq-vae. *Proc. CVPR*.

Ren, Y., Yu, X., Zhang, R., Li, T. H., Liu, S., Li, G., 2019. Structureflow: Image inpainting via structure-aware appearance flow. *Proc. ICCV*.