Improving True Ortho Quality by Removing Moving Vehicles

HANNES NÜBEL[1] & PATRICK TUTZAUER[2]

Abstract: Geometrically accurate representations like True Orthophotos (TOs) and textured meshes generated from aerial imagery are utilized progressively in many fields. Because of the growing interest in such products, these are optimized on many fronts like geometry, texturing, or processing time. An important customer need regarding the TO is the appropriate handling of moving vehicles. Apart from the ability to remove such objects completely, it is especially desired to prevent possible artifacts that are caused by a blending of varying textures from different images – an effect referred to as ghost cars. This paper describes an approach to remove moving vehicles from TOs by masking them in the aerial imagery with help from the information of depth images using a Convolutional Neural Network (CNN). Those masks are then deployed for the texturing of the TO. The approach proved to be effective with ~90% of the moving vehicles detected within the testing images and results in TOs with drastically removed moving vehicles and artifacts.

1 Introduction

Capturing aerial imagery is a time-efficient method to map residential areas in a short time frame. However, because of the perspective projection of the image and the varying heights of the captured surface, they do not represent a map-like regular grid on the earth's surface. By means of a standard photogrammetric workflow, multiple overlapping images, initial positions, and orientations can be utilized to compute and refine 3D points and orientations, performing a bundle adjustment (KRAUS 2011). That information can then be used to map the texture information from the images onto the specified regular grid, creating an orthophoto. To be able to calculate the depth of corresponding points, they must be visible in at least two images.

One assumption in this process is that all captured objects are steady while taking the series of images. While geometric reconstruction from multiple views works reliably in most parts, texturing can pose difficulties. To get a consistent texture in the resulting product, it is common to use the color information out of multiple images for a single pixel in the orthophoto. If the information is taken from only one image, the results can get patchy, because of the different lighting conditions and viewing angles in the respective images. If the color of a pixel in the orthophoto is computed as the mean of the contributing pixels in multiple images, it is possible that the texture of a moving object - which is never at the very same position in multiple images - is blended with the surface. This can lead to effects like the so-called ghost cars shown in Fig. 1.

This paper demonstrates the methodology and results of an approach that is able to remove such artifacts by detecting moving vehicles in the input imagery and masking them during the generation of the TO.

[1] Universität Stuttgart, Institut für Photogrammetrie, Geschwister-Scholl-Straße 24D, D-70174 Stuttgart, E-Mail: hannue@gmx.de

[2] nFrames | Esri R&D Center Stuttgart, Kornbergstraße 36, D-70176 Stuttgart, E-Mail: ptutzauer@esri.com

 DOI: 10.24407/KXP:1841069671

Fig. 1: Ghost cars induced by blending the texture of vehicles and street from different aerial images.

2 Related Work

One of the main scopes of the proposed approach is the detection of cars in images. This task has been discussed in many scientific papers and is now frequently implemented in the form of CNNs. For car detection, it is often required to detect each as a single object (e.g to derive the number of cars). Therefore, instead of semantic segmentation, where each pixel in an image obtains a label, this work relies on instance segmentation to derive a mask and class label for each detected object. AUDEBERT et al. (2017) are using a three-step approach to first perform a semantic segmentation, then derive vehicle objects by searching for corresponding connected pixels, and finally, classify the instances into different types of vehicles. In the same year, Mask R-CNN (HE et al. 2017) was published, which was presenting an approach for a single CNN performing instance segmentation with masks on a pixel level. Mask R-CNN is also used as a base to solve the vehicle detection task in this study.

In contrast to other studies, this approach is utilizing aerial imagery together with photogrammetric data in the form of depth images, to only detect moving vehicles. The position and extent of each detected object are used during the texturing of the TO to exclude those image regions, which otherwise could lead to artifacts caused by the blending of inconsistent textures. Furthermore, it prevents texture in the TO that does not represent the reconstructed geometry in the Digital Surface Model (DSM).

3 Detection and Masking of Moving Vehicles

To be able to exclude image regions that show moving vehicles from the TO generation, it is necessary to create masks that indicate which pixels represent moving vehicles and should therefore be ignored. The mask for an aerial image is derived using an adapted and trained version of the Mask R-CNN (Fig. 2), which is performing instance segmentation to derive a semantic mask for each moving vehicle in the corresponding image. The CNN consists of a backbone network, which produces feature masks. To identify areas that could contain the desired objects, anchor boxes are learned using a Region Proposal Network. The pixels from the feature map of those anchor boxes are then passed on to determine the class, as well as a pixel-accurate mask of the respective object within the image section. Since aerial images are typically taken by large format cameras, and therefore are very large, they get tiled before being fed into the adapted Mask R-CNN. This also allows to keep a consistent input image size without resampling and stretching the original image.

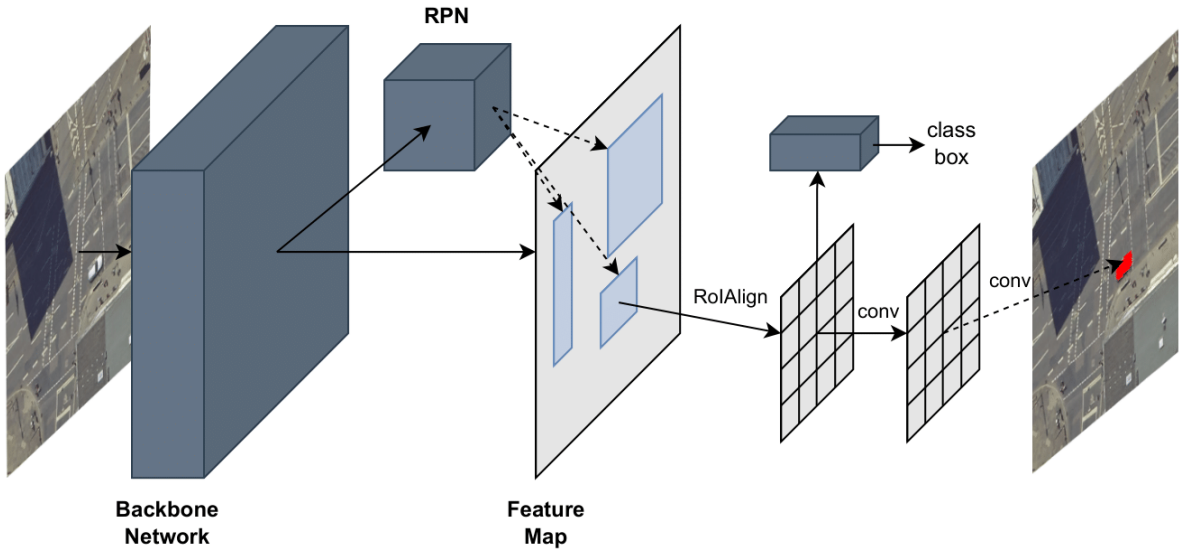

Fig. 2: Mask R-CNN framework (adapted from HE et al. 2017).

In addition to the color images captured by the camera, the respective depth images are used as a second input to the CNN. Although it is possible to check if certain objects are stationary in a series of images, this information is not contained in a single color image. The reconstruction pipeline SURE (ROTHERMEL et al. 2012) is producing depth images as an intermediate product from which such information about stationariness can be derived implicitly. A depth image depicts the distance of the ground points to the camera for the matched pixels in the base image. However, these depth images cannot be completely filled with values. Depth values can only be derived in areas where the used images are overlapping, so that corresponding pixels in stereo image pairs can be matched. Furthermore, for some image areas it is not possible to derive depth values, because the respective areas in the object space are occluded for the stereo image. Pixels of objects that are moving between the capturing of images also cannot be matched, because the pixels in this location show a different object than the pixels of another image mapping the same region. In

the case of moving cars, this means that an image that contains a car at a certain position will show the surface of the road for the same region in a different image. This results in holes in the depth image when a moving vehicle is captured. Moving vehicles can therefore be detected for each image when using the color information from an RGB image in conjunction with the depth image to tell if they are stationary (Fig. 3). The image stack consisting of RGB and depth image is fed to the CNN, which produces masks for the moving vehicles. Those are then used in the TO texturing process, to exclude the pixels in the color images and use color information from different views instead.

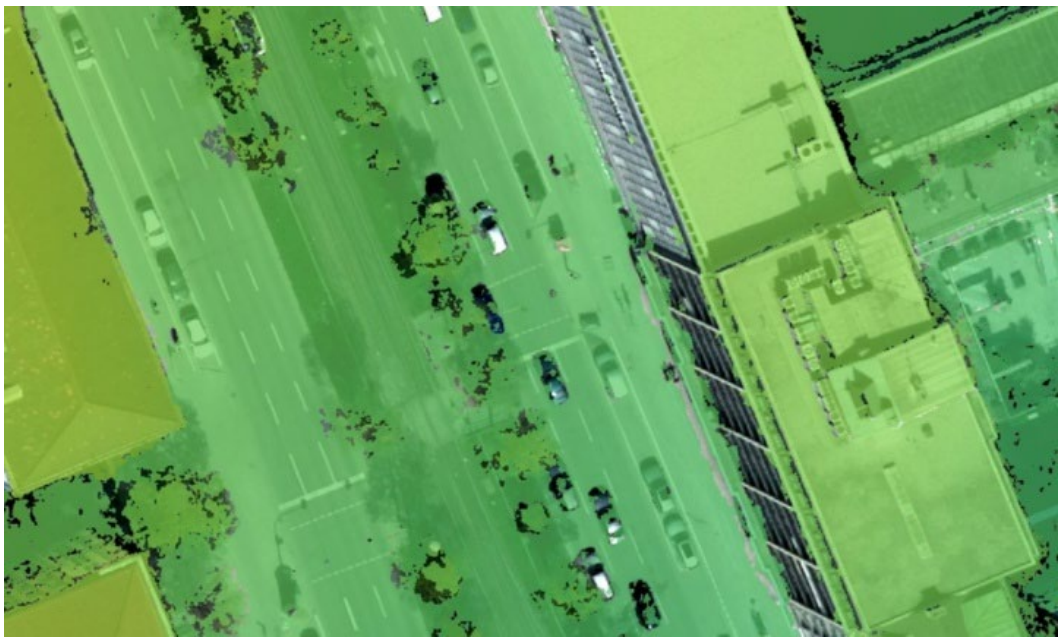

Fig. 3: RGB with transparent overlaid color-coded depth image (NaN values for depth image shown as no data) showing the difference between moving and stationary vehicles in the depth image.

To be able to train the CNN to identify moving vehicles, a large amount of training data is necessary. To create ground truth data for all aerial images, an existing car detection CNN was used to create masks for all cars, relying only on RGB information. To generate as many detections as possible, the probability threshold was set very low. The resulting detections were then automatically reduced by the cars that were clearly stationary because of a complete coverage of depth values. While strongly reducing manual labor, it was still necessary to delete false positives and add undiscovered moving vehicles by hand. Of the thereby created data 80% was used for training, while 20% was held out for final testing evaluations.

The process to create an improved TO without ghost cars was implemented as an extension to the reconstruction software SURE in a proof-of-concept fashion. It can be executed as one pipeline and does not need further manual interference, apart from setting up the standard inputs and parameters. After SURE has completed the dense matching and written out the depth maps for each base image, those are used together with the input imagery as input for the adapted Mask R-CNN, which is predicting pixel-accurate masks for all moving vehicles. When the DSM was processed, the moving vehicle masks for each image are used as a secondary input to the TO texturing process (Fig. 4).

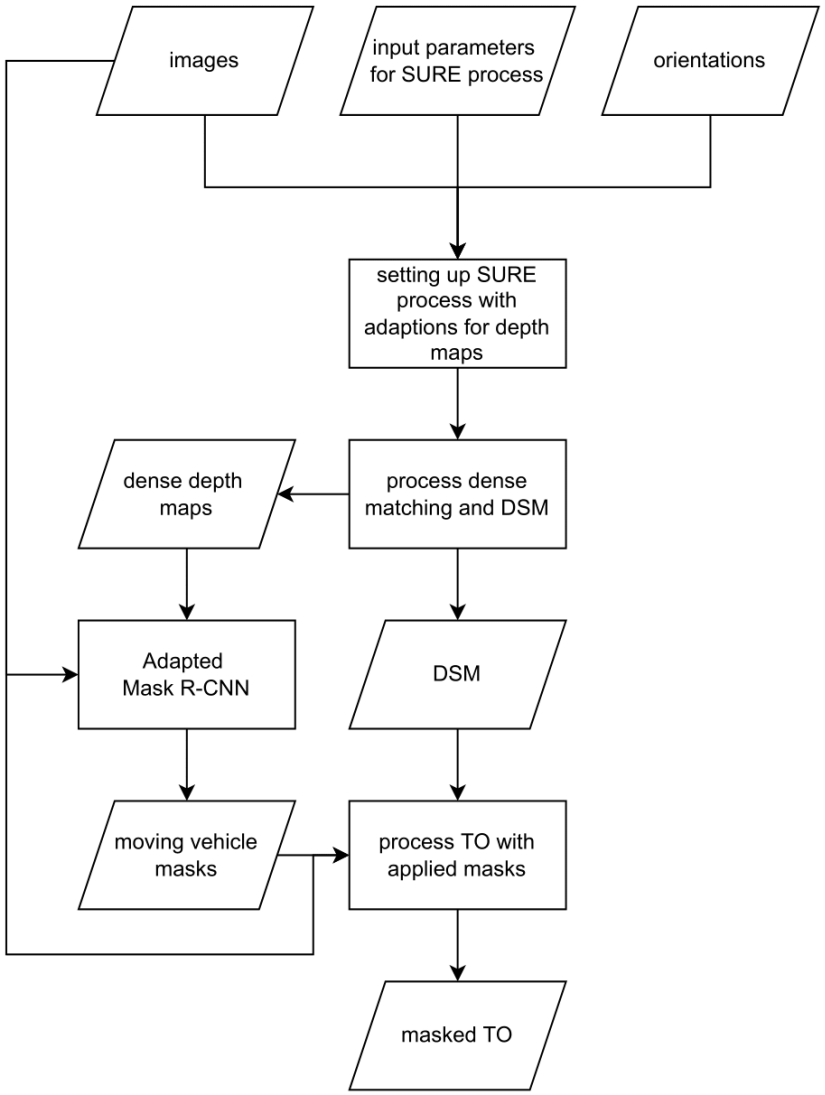

Fig. 4: Flowchart of pipeline integrating Mask R-CNN and SURE.

4 Results

4.1 Aerial Images with Moving Vehicle Masks

Fig. 5 shows the successful detection of cars and indicates that the CNN is also able to distinguish between moving and stationary. It also highlights a very challenging issue when vehicles are not constantly and/or only slowly moving. In this case, some points on the vehicle can be matched between images, resulting in an inconsistent set of depth values, as can be seen for the cars coming from the bottom and turning right. This is not only difficult when trying to predict an unseen image, but already when creating the ground truth for the training data. It has to be decided based on color and depth image, whether a vehicle is moving or not. If potentially moving objects are excluded from the ground truth masks, those could lead to unwanted effects in the TO if such vehicles do not get detected. On the other hand, if they are included in the training, many false positives might be detected for data sets or areas with sparsely filled depth images.

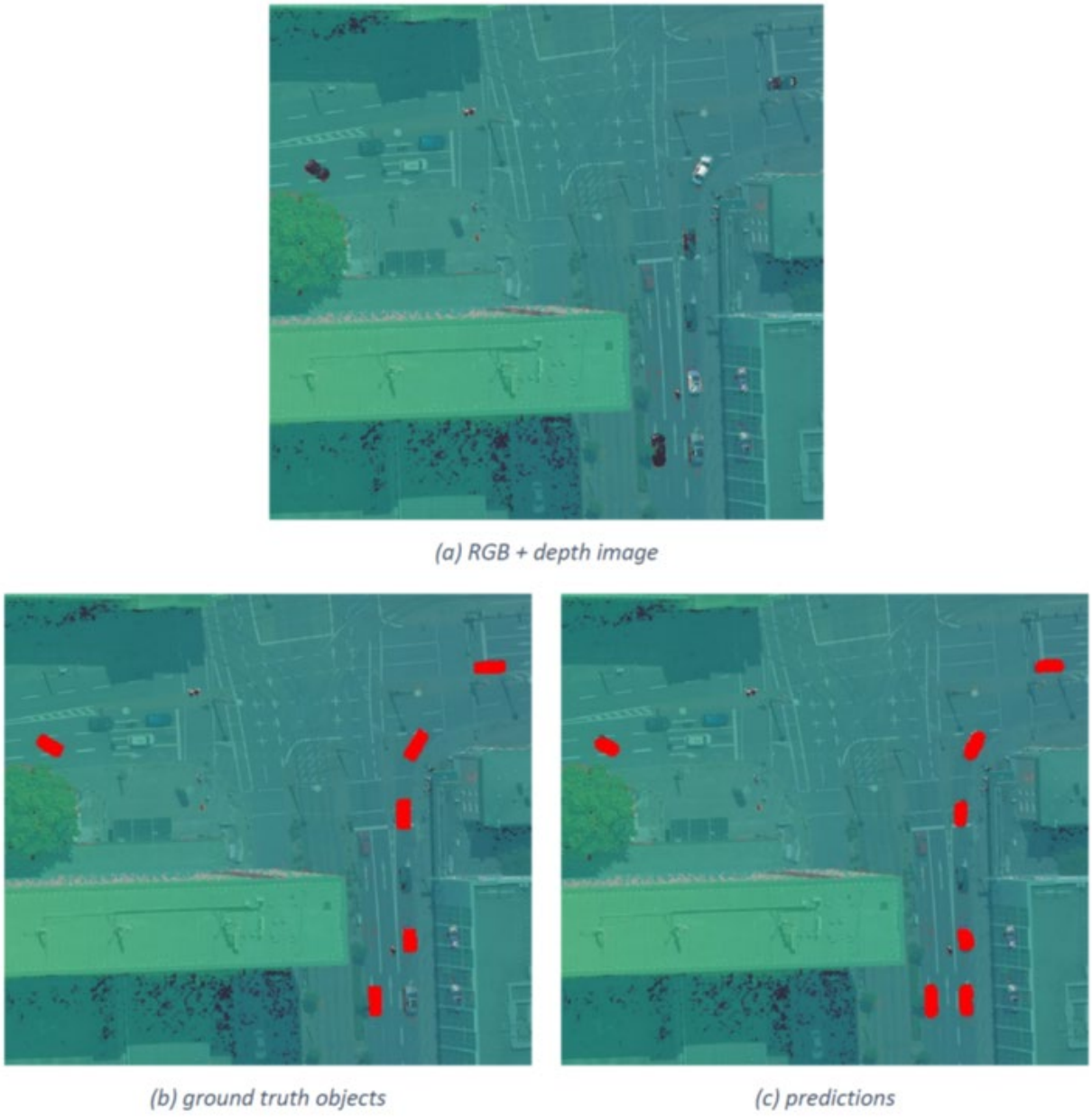

Fig. 5: Comparison of ground truth and prediction with moving, stationary, and partly moving vehicles.

If depth images have many invalid pixels, the precision of the CNN prediction drops because it is not possible to distinguish areas with moving objects from areas with insufficient matching caused by other issues like low overlap in the depth images. Since moving objects are one of the main contributors to missing depth information, completeness of depth maps is preferred over highest possible accuracy in the parameterization of this work.

4.2 Masked True Orthophotos

Fig. 6 shows segments of a TO with moving vehicles and the standard texturing approach on the left. The masked version, excluding the corresponding pixels in the aerial images from the texturing of the TO, is shown on the right. It can be observed that moving vehicles are effectively removed from the TO while the stationary ones, at traffic lights or parking spaces, retain their visualization. The approach is removing moving vehicles that are displayed completely in the reference, as well as ones that are blended, cut, textured multiple times and combinations of such effects that are especially unpleasant.

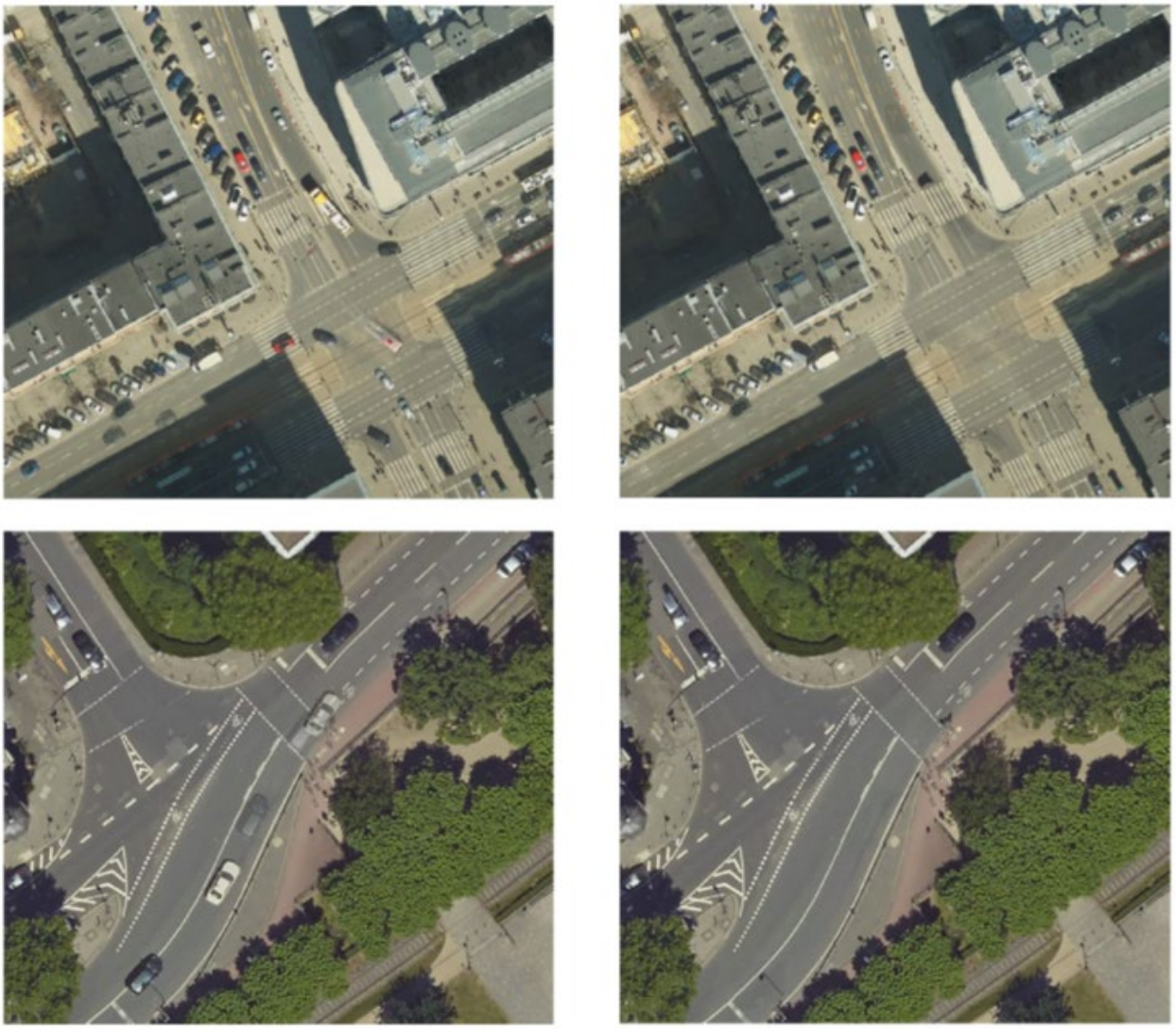

Fig. 6: TO comparison reference (left) and masked (right).

While detecting many moving vehicles in aerial images, the approach still relies on sufficient texture for those regions in other images. If the images have different lighting conditions, or never show the surface of the road (e.g. because of extreme traffic), data gaps or inconsistent texture for the masked areas in the TO can occur.

Figure 7 shows another data set with a very crowded road segment. The approach proves to be very effective in such challenging scenarios as well. All of the flowing traffic gets removed and on the left side replaced by matching textures from different images. On the right, there are some artifacts in the areas where cars got removed. This occurs because some of the images were captured with a significant temporal shift, causing the shadows of the larger buildings to move. Having no texture information for the areas of the moving vehicles, the texturing algorithm has to rely on texture information from other images, which in this case is inconsistent with the lighting conditions of their surroundings.

Fig. 7: Comparison of reference (top) and masked (bottom) TO with varying lighting conditions for different images on the right of the excerpt.

5 Conclusion

The approach described in this work proved to fulfill the need of removing moving vehicles from TOs for a high percentage of objects, creating a better view of the road networks, as well as mitigating many undesirable effects like ghost cars coming from an inconsistent blending of textures from moving vehicles. Since this approach is not changing the actual texturing procedure, the illustration of areas that do not have to be masked is staying the same. In addition, areas with masked moving vehicles mostly blend in with their surroundings since the texturing is performed equally by simply excluding the specified area from an image. However, the similarity is of course dependent on the views available for the corresponding area.

Two main challenges were encountered with the proposed approach. One is the necessity for dense depth maps to be able to reliably distinguish moving from stationary vehicles. The second is the handling of vehicles that cannot be labeled explicitly as moving or stationary for all images in which they are captured. This may lead to vehicles not being removed, i.e. appearing as for the reference TO, although they might cause artifacts. Nevertheless, the majority of moving vehicles are being detected and removed especially for flowing traffic, creating a more appealing end result.

To further improve the precision of the CNN it would be possible to include other byproducts from the reconstruction pipeline that provide additional information about the matching of points. Apart from that, the detections could be classified into different kinds of moving vehicles, to be able to reliably detect vehicles of different shapes and sizes.

6 References

AUDEBERT, N., LE SAUX, B. & LEFÈVRE, S., 2017: Segment-before-detect: Vehicle detection and classification through semantic segmentation of aerial images. Remote Sensing, **9**(4), 368.

GOODFELLOW, I., BENGIO, Y. & COURVILLE, A., 2016: Deep Learning. MIT press.

HE, K., GKIOXARI, G., DOLLÁR, P. & GIRSHICK, R., 2017: Mask r-cnn. Proceedings of the IEEE international conference on computer vision, 2961-2969.

KRAUS, K., 2007: Photogrammetry: Geometry from Images and Laser Scans. Berlin, Boston: De Gruyter, https://doi.org/10.1515/9783110892871.

REN, S., HE, K., GIRSHICK, R. & SUN, J., 2015: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems, **28**.

ROTHERMEL, M., WENZEL, K., FRITSCH, D. & HAALA, N., 2012: SURE: Photogrammetric surface reconstruction from imagery. Proceedings LC3D Workshop, Berlin.