OBJECT DETECTION FROM MMS IMAGERY USING DEEP LEARNING FOR GENERATION OF ROAD ORTHOPHOTOS

Y. Li [1,*], M. Sakamoto [1], T. Shinohara [1], T. Satoh [1]
[1] PASCO CORPORATION, 2-8-10 Higashiyama, Meguro-ku, Tokyo 153-0043, Japan -
(yiorn_3951, moittu9191, taarkh6651, tuoost7017)@pasco.co.jp

Commission II, WG II/4

KEY WORDS: Object detection, Vehicle and its shadow detection, Deep learning, Faster R-CNN, Road orthophoto, Mobile mapping system

ABSTRACT:

In recent years, extensive research has been conducted to automatically generate high-accuracy and high-precision road orthophotos using images and laser point cloud data acquired from a mobile mapping system (MMS). However, it is necessary to mask out non-road objects such as vehicles, bicycles, pedestrians and their shadows in MMS images in order to eliminate erroneous textures from the road orthophoto. Hence, we proposed a novel vehicle and its shadow detection model based on Faster R-CNN for automatically and accurately detecting the regions of vehicles and their shadows from MMS images. The experimental results show that the maximum recall of the proposed model was high—0.963 (intersection-over-union>0.7) —and the model could identify the regions of vehicles and their shadows accurately and robustly from MMS images, even when they contain varied vehicles, different shadow directions, and partial occlusions. Furthermore, it was confirmed that the quality of road orthophoto generated using vehicle and its shadow masks was significantly improved as compared to those generated using no masks or using vehicle masks only.

1. INTRODUCTION

In recent years, with the widespread application of the mobile mapping system (MMS), it has become possible to collect images and point cloud data efficiently on road and surrounding environment. High-definition road orthophoto is one of the most important products generated using images and point cloud data acquired by MMSs. In contrast to aerial surveys, an MMS acquires images and point cloud data from a position that is very close to the road surface; therefore it is possible to generate a very high-definition road orthophoto, e.g., a road orthophoto with 1-cm resolution. Road orthophotos have been widely applied in various fields. For example, we use road orthophotos as the background of digital maps, or extract road edges, road signs or road cracks from a road orthophoto.

A method for generation of road orthophotos using images and point cloud data acquired from MMSs has already been proposed in our previous research (Sakamoto et al., 2012). Firstly, the point cloud of the road surface are extracted from the point cloud data by applying a filtering technique. A 3D triangulated irregular network (TIN) model is then generated using the road surface point cloud to represent the topography of the road surface. Finally, the RGB values at each pixel of the road orthophoto are calculated using a texturing technique. Furthermore, in our subsequent study (Li et al., 2017), erroneous vehicle textures have been efficiently removed from road orthophotos using vehicle masks, which are detected automatically from MMS images by Faster R-CNN. However, the shadows of the vehicles are left as black spots in the road orthophoto, which is undesirable.

We proposed a novel detection model for vehicle and its shadow (VaS) to accurately and automatically detect both of the VaS regions from MMS images for the purpose of removing erroneous textures that appear in road orthophotos, which are generated using images and point cloud data acquired from MMSs.

2. METHODOLOGY

In this chapter, we first describe how to generate a road orthophoto using VaS masks, then analyse the features of a VaS detection model suitable for the above application, and finally explain our proposed VaS detection model.

2.1 Road Orthophoto Generation Using VaS Masks

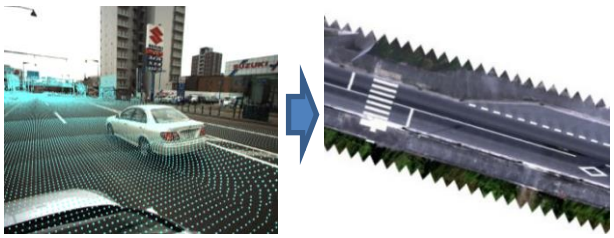
Figure 2 illustrates the difference between two texturing processes, with and without VaS masks. In the former process, the RGB values of a pixel in the road orthophoto are generally obtained from the image, which has the highest resolution at the corresponding position. In contrast, in the latter process, the textures are obtained only from the areas excluding the masks, in order to generate a high-quality road orthophoto with no erroneous textures.

2.2 Features of VaS Detection for Road Orthophoto Generation

As shown in Figure 3, two methods are available for VaS detection: (1) the detection of the region of VaS separately, and (2) the detection of the region that includes both of the VaS. We select the latter method for the reasons mentioned below.

In this study, the shadows that are required to be detected are only of moving objects such as vehicles. This is because, if the shadow regions of fixed objects such as buildings and poles are used as masks in the texturing process, the corresponding areas in the road orthophoto become occlusions. However, in the former method, it seems to be difficult to distinguish the shadows of moving objects from those of fixed objects. Hence, we select the latter method that involves the detection of the VaS as one region. Furthermore, the latter method has an advantage in the preparation of the ground truth. Because we need to create only one bounding box for each VaS.

* Corresponding author



Images and point clouds High-definition Road orthophoto

Figure 1. Generation of high-definition road orthophoto using images and point clouds acquired from MMSs

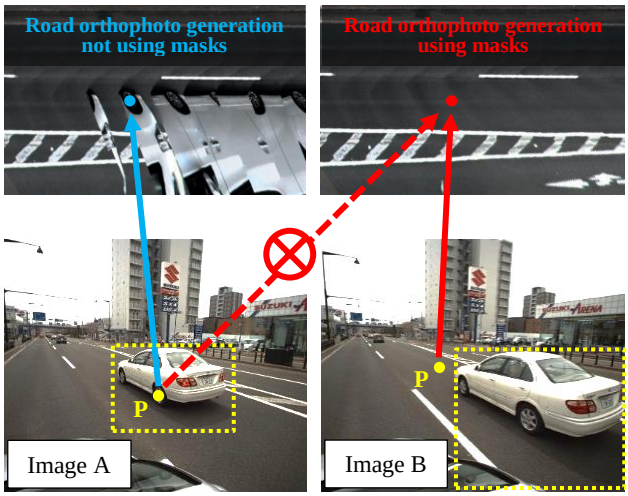
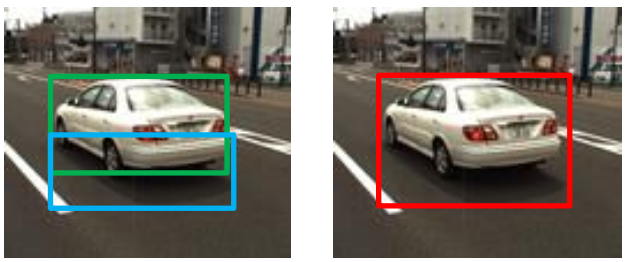


Figure 2. Road orthophoto generation without and with the use of VaS masks



(1) Detect VaS separately (2) Detect VaS as one region

Figure 3. Two available methods for detecting VaS

The requirements for object detection depend on applications. In the case of this study, the objects to be detected are VaSs located within a certain distance from the camera, e.g., 10 m. In order to obtain a high-resolution road orthophoto, it is necessary to acquire textures from the region of the images having a resolution that is equal to or higher than those of the road orthophoto. Hence, it is unnecessary to detect small VaSs located in the regions that are far from the camera. It is also necessary to accurately detect the VaS regions. Part of the VaS region that are not included in the detection region are at a risk of being used as textures of the road orthophoto.

2.3 Proposed VaS Detection Model

Based on the analysis in the previous section, we propose a highly accurate VaS detection model based on Faster R-CNN.

The existing object-detection methods include one-stage methods, such as You Only Look Once (YOLO) (Redmon et al., 2016), Single Shot MultiBox Detector (SSD) (Liu et al., 2016) and two-stage methods such as Fast R-CNN (Girshick, R., 2015) and Faster R-CNN (Ren et al., 2015). The former detects objects directly in an input image, and the latter first roughly proposes candidates from an input image, and then corrects the classification and region of the candidates to obtain an accurate result. In general, the former has a high processing speed while the latter has a high detection accuracy. The generation of road orthophotos generally occurs at the post-processing stage, and thus, there is no critical requirement for a high processing speed. Hence, we select Faster R-CNN as the base model for VaS detection to achieve high accuracy of detection of the VaS region.

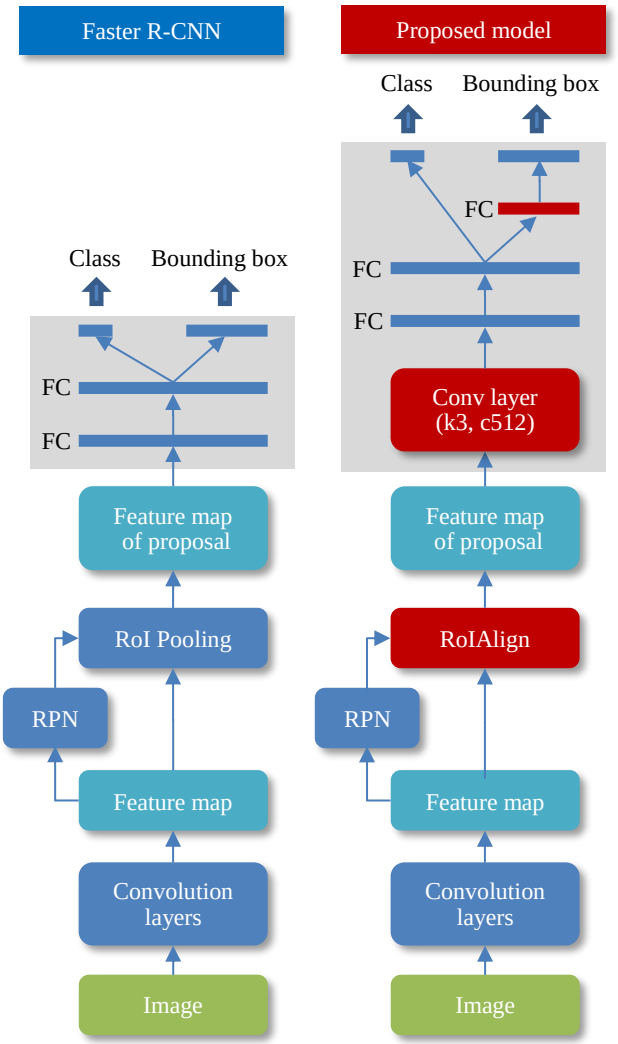


Figure 4. Faster R-CNN and proposed model

As shown on the left side of Figure 4, Faster R-CNN first performs several convolution and pooling processes on the input image to generate a feature map. Next, candidate regions are proposed using region proposal network (RPN) on the feature map. Finally, after pooling the range, which are corresponded to the candidate region, from the feature map by region of interest (RoI) pooling, the processes of classification and region correction are performed based on the feature map of the proposal through two fully connected (FC) layers. The details may be referred to in the original paper.

We propose a novel VaS detection model (right side of Figure 4) based on Faster R-CNN in order to improve the accuracy of detection and especially the accuracy of bounding box.

2.3.1 RoIAlign

As described above, in the RoI pooling, processing is performed to extract the range corresponding to the candidate region from the feature map to the predetermined size. However, a pooling is performed for integer pixels in the RoI pooling, when the range of the candidate region on the feature map is not an integer pixel, an adjustment is made in order to obtain an integer pixel for that range. Owing to this adjustment in the RoI pooling, a deviation occurs between the candidate region and the actual feature extraction range, which may result in a lowering of the accuracy of the detection region.

To solve the above problem, He et al. (2017) have proposed RoIAlign. While pooling processing have to be performed in integer pixel units in RoI pooling, the pooling process is extended to allow it to cope with a subpixel unit by using an interpolation method in RoIAlign. As a result, RoIAlign completely matches the candidate region and the extraction range in the feature map. In this research, we adopt RoIAlign instead of RoI pooling for the purpose of improving the accuracy of detection of the VaS regions.

2.3.2 Enhancement of Network Performance with Respect to the Bounding Box

According to the research by Ren et al. (2016), the network of the second stage (the network of the grey background in Figure 4) greatly affects the detection accuracy. Therefore, in this research, the following two improvements were made for enhancing the performance of the network.

① A convolution layer with a kernel size of 3 × 3 pixels and 512 channels was added before the first FC layer.
② An FC layer with 1024 nodes was added before outputting the bounding box.

The purpose of the first improvement is to extract more appropriate features and that of the second is to enhance the performance of the network with respect to the bounding box.

2.3.3 Weight of Losses

In the training phase, it was possible to obtain ideal results by adjusting the weight of the losses. We assign more weight to the loss of the bounding box than those of others in order to improve the accuracy of the bounding box rather than that of the classification.

3. EXPERIMENTAL RESULTS AND DISCUSSION

3.1 Experimental Data

Images and point cloud data were obtained from MMS by traveling on four courses over a total of 2.4 km for experiments. The size of the images was 2,400 × 2,000 pixels, and the resolution after 10 m was approximately 1 cm. In addition, the images were captured at intervals of 2 m. Moreover, the density of the point cloud, which was acquired using a 2D laser scanner manufactured by SICK Corporation, was approximately 100 points/m^2 on the road surface.

Regarding the training data, 100 consecutive images were selected from each course, and a total of 400 images were prepared for the training. In order to increase the number of training images, nine types of images were created by adjusting the saturation (S) and value (V) to 0.7, 1.0 and 1.5 times for the original images in the HSV color space. Hence, 3,600 images were used in the training process. Furthermore, 140 images were used as the test data, which included 100 images selected from the above experimental images and 40 images selected from different experiment data. The latter test images were used to evaluate the generalization performance of the trained model. The ground truths were created manually and targeted at VaSs within 10 m from the camera to generate a road orthophoto with a 1-cm resolution.

3.2 VaS Detection from MMS Images

First, the proposed model of VaS detection was trained using 3600 training images. In the case of the input images, the multi-scale images were utilized at the training stage. Specifically, while maintaining the aspect ratio of the original image, the length of the short side was randomly set as either 600, 800, 1000, 1200 or 1500 pixels each time. Furthermore, in the inference stage, fixed size images of 1,200 × 1,000 pixels were used. The weight of the bounding box loss was set as 4.0, while the weight of the other losses were set as 1.0. The hyper-parameters for the training were set as shown in Table 1, and Caffe (Jia et al., 2014) is adopted as the framework. In addition, we adopted transfer learning for efficient training, and the object detection model, which has been trained using MS COCO data set (http://cocodataset.org), was used as a pre-trained model. As shown in Figure 5, the training loss decreased steadily and eventually converged to almost 0.1.

Items	values
Optimization	Stochastic Gradient Descent
Learning rate	0.001 (set as 1/10 per 20,000 iterations)
Maximum iteration	60,000
Momentum	0.9
Weight decay	0.0005

Table 1. Hyper-parameter settings

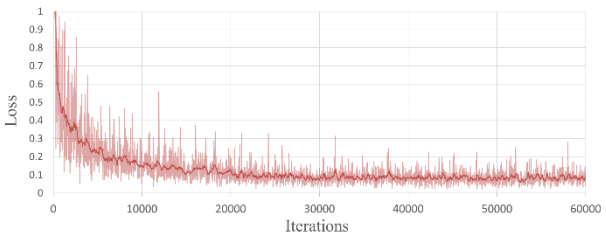


Figure 5. Training loss

Next, we applied the trained model to 140 test images to detect VaS regions, and then verified the accuracy by comparing the obtained results with the ground truth. The result of detection was regarded as a success only if the intersection-over-union (IoU) between the region of detection and those of the corresponding ground truth was over 0.7. As shown in Figure 6, the average precision (AP) was 0.895, breakeven point was 0.828, and maximum recall was 0.963, which show that very high accuracy is realized.

However, it should be noted that the precision is underestimated As described above, as the detection targets are the VaSs within 10 m from the camera, when VaSs that are far from the camera are detected, they are considered as false positives (Figure 7). Therefore, for models with a higher performance that can detect

small VaSs in the distance, a lower evaluated precision is obtained. Similarly, the AP and breakeven point are also underestimated for these models. Therefore, in this research, the maximum recall is used as an evaluation index. In fact, in the detection results of 140 test images obtained using the proposed model, there were only six false positives outside the VaS areas. Hence, in the application to the generation of a road orthophoto—as in this research—the maximum recall is considered to be appropriate as the evaluation index of the detection result.

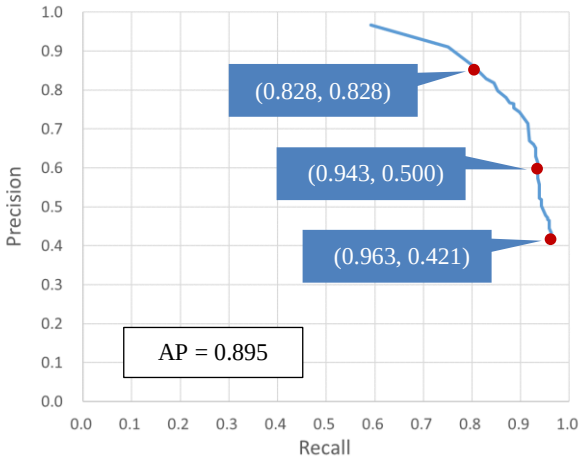


Figure 6. Precision-Recall curves (IoU > 0.7)

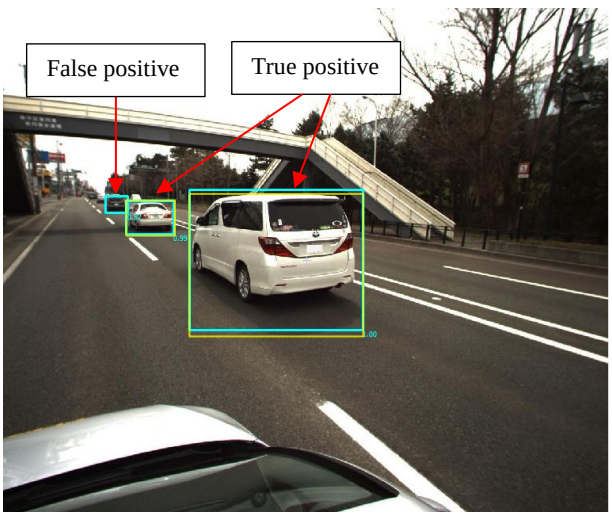


Figure 7. Results of VaS detection obtained using proposed model (cyan box: detection results, and yellow box: ground truth)

Figure 8 shows some representative examples of the VaS detection results obtained using the proposed model. From those successful cases (blue-bordered images in Figure 8), it was confirmed that the proposed model can accurately detect the VaS regions from MMS images without being affected by the type of vehicles, shadow directions, partial occlusions, etc. Further, even in an extreme case wherein only a very small part of a VaS has appeared, the proposed model was found to be robust in detection. Furthermore, from examples of the failure cases (red-bordered images in Figure 8), it can be observed that there is a misdetection of a large track, over-detection of a motorbike, and an inaccurate detection region, which must be addressed in future research.

3.3 Generation of Road Orthophoto

A road orthophoto was generated using VaS masks created based on the VaS region detected using our proposed model. Only the detection results with a probability greater than 0.5 were used. For comparison, a road orthophoto without the use of a mask and a road orthophoto with the use of vehicle masks were generated.

Firstly, as shown in the upper part of Figure 9, it was confirmed that parts of the vehicles were used as a texture in the road orthophoto that was generated without the use of a mask. Moreover, in the road orthophoto generated using vehicle masks, the vehicle textures that appeared in the above road orthophoto were eliminated but the shadows of the vehicle were retained as black spots (middle part of Figure 9). In the case of the use of the VaS masks (lower part of Figure 9), it was confirmed that the erroneous textures were completely removed from the road orthophoto. The comparison shows that the proposed method is extremely effective in the generation of a high-quality road orthophoto.

4. CONCLUSION

A novel VaS detection model was proposed for eliminating the erroneous textures from the high-definition road orthophoto generated using images and point cloud data acquired through MMSs. In the proposed model, we attempted to improve the accuracy of detection, especially the accuracy of detection of the VaS regions, through the introduction of several improvements based on Faster R-CNN. For the purpose of evaluation, the proposed model was applied to the experimental images to detect VaSs from MMS images. A high-definition road orthophoto was then generated using the VaS masks, which were created from the regions detected using our proposed model automatically. The experimental results show that the maximum recall of the proposed model was very high—0.963 (IoU > 0.7) and also very robust against variations in the vehicle types, shadow directions, and existence of partial occlusions. Furthermore, it was confirmed that the quality of the road orthophoto generated using VaS masks was significantly better than the road orthophoto generated using vehicle masks only.

In future works, bikes, bicycles, pedestrians etc. are also required to be detected in MMS images in addition to vehicles for the purpose of eliminating erroneous textures from a road orthophoto. Moreover, it is necessary to detect the exact region of the target objects in order to reduce the area of the road surface included in the detected regions. As such over-detected regions of the road surface could lower the resolution of the obtained road orthophoto or cause the occurrence of occlusions in extreme cases.

REFERENCES

Girshick, R., 2015. Fast R-CNN, ICCV, pp.1440-1448.

He, K., Gkioxari, G., Doll´ar, P. and Girshick, R., 2017. Mask RCNN. *ICCV*.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. *CVPR*.

Li, Y., Sakamoto, M., Shinohara, T. and Satoh, T., 2017. Object Detection from MMS Imagery Using Deep Learning for Generation of Road Orthophotos, *ISPRS Techinical Commision II Symposium*.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., and Berg A. C., 2016, SSD: Single Shot MultiBox Detector, *ECCV*.

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A., 2016. You Only Look Once: Unified, Real-Time Object Detection. *CVPR*.

Ren, S., He, K., Girshick, R., and Sun, J., 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *NIPS*, pp. 91-99.

Ren, S., He, K., Girshick, R., Zhang, X. and Sun., 2016. Object detection networks on convolutional feature maps. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sakamoto, M., Tachibana, K., and Shimamura, H., 2012. Generation of High Resolution and High Precision Orthorectified Road Imagery from Mobile Mapping System. *ISPRS*, 39(B5), pp. 499-504.

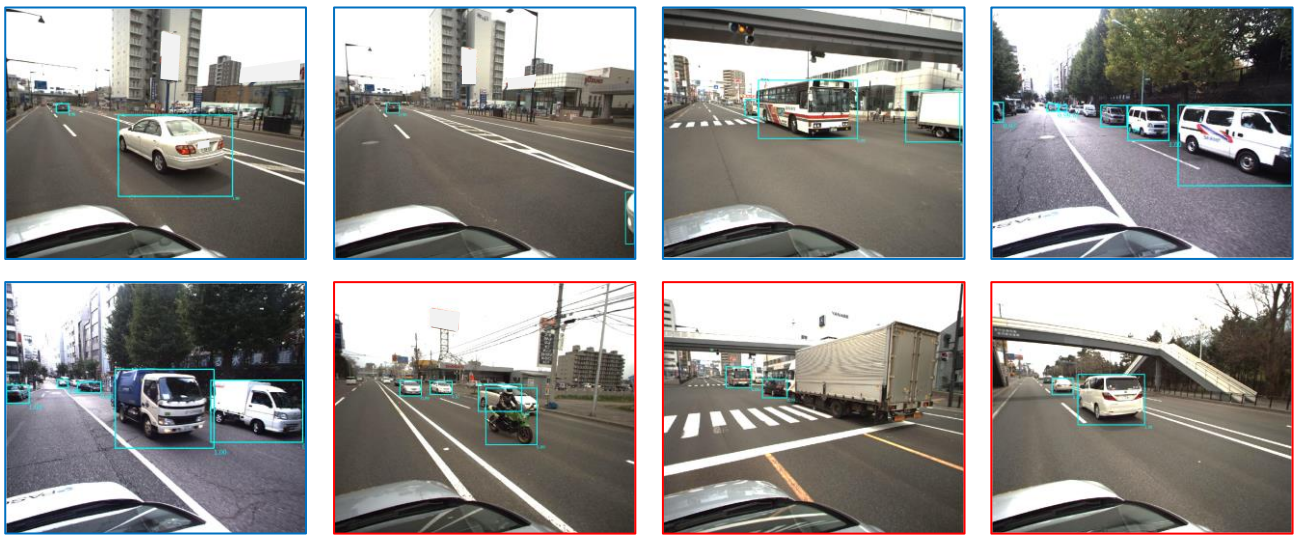


Figure 8. Results of VaS detection obtained using our proposed model

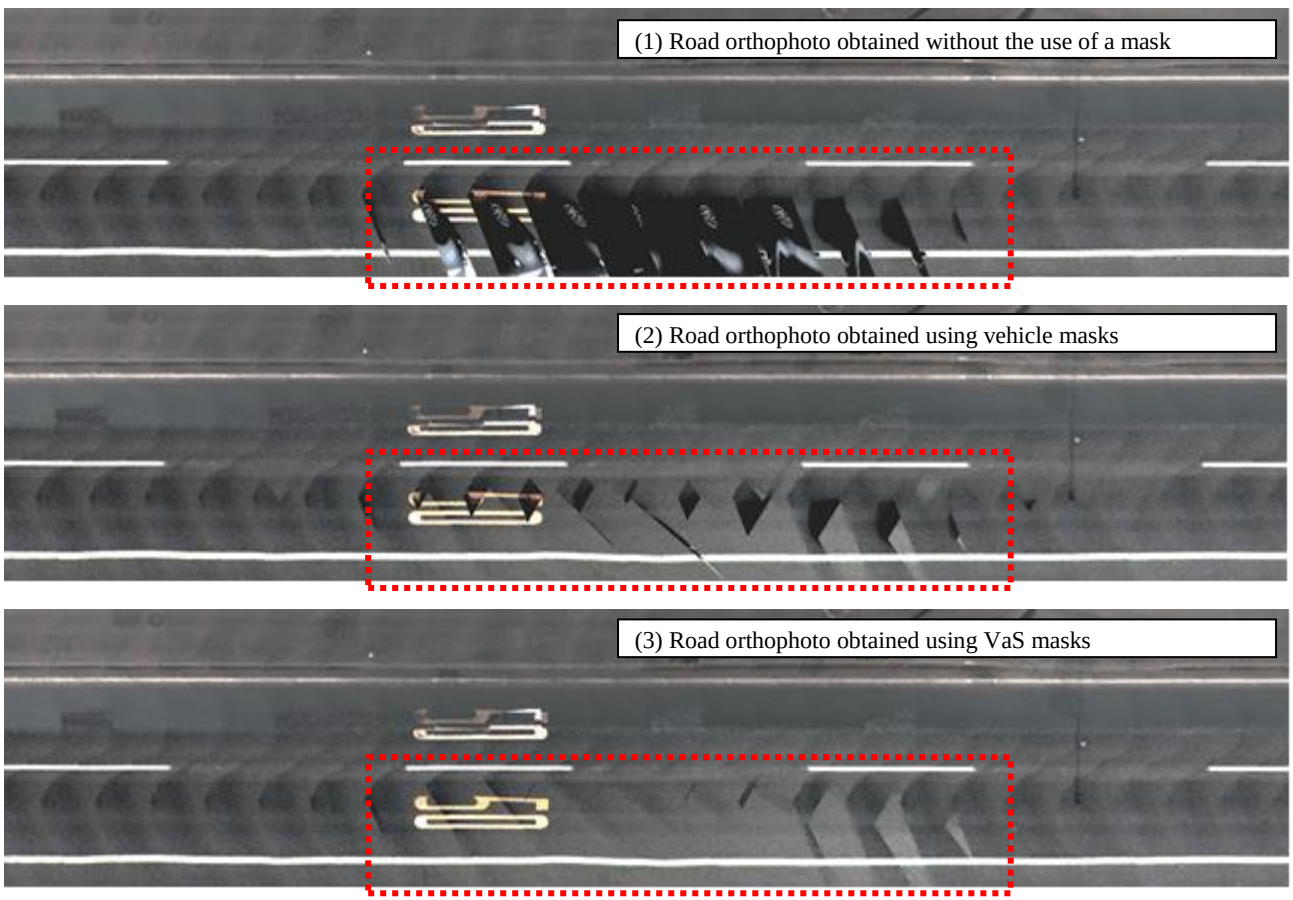


Figure 9. Comparison of road orthophotos obtained (1) without the use of a mask, (2) using vehicle masks, and (3) using VaS masks