Supplementary Material: Orientation-aware Vehicle Re-identification with Semantics-guided Part Attention Network

Tsai-Shien Chen[1,2], Chih-Ting Liu[1,2], Chih-Wei Wu[1,2], and Shao-Yi Chien[1,2]

[1] Graduate Institute of Electronic Engineering, National Taiwan University
[2] NTU IoX Center, National Taiwan University
{tschen, jackieliu, cwwu}@media.ee.ntu.edu.tw
sychien@ntu.edu.tw

1 Details of Generating Foreground Vehicle Masks

To get the foreground mask of the whole vehicle, we use a traditional segmentation technique, Grabcut [4]. However, it requires user to frame the target object out from the whole image for the first stage segmentation, and mark a part of wrong-labeled pixels for obtaining a better result. Yet, neither of them can be done manually owing to the large scale of our dataset. Considering that the input vehicle images in our dataset are all first generated by vehicle detection algorithm, we utilize an automatic method that assumes the pixels on the image border all belong to the background, and therefore we can frame out the object from the border-padding image with the original image size to get the first stage segmentation result.

To acquire more robust background-removed image, we use the first stage results as target labels to train a segmentation CNN network with one ResNet-50 [1] followed by four transposed convolutional layers. But, to avoid the network overfitting on the unstable results generated by Grabcut, after training for a few epochs, we remove the training images with abnormal huge loss, which possibly represent unsatisfactory results done by Grabcut. Finally, we use this trained segmentation CNN network to inference all the data to get the background-removed images.

As shown in Fig. 1, we visualize some unfavorable background-removed images generated by Grabcut. The first stage results are unsteady; the background region is sometimes mistakenly segmented to foreground while some parts of vehicle which may contain the discriminative features such as wheels and headlamps are sometimes classified to background. In contrast, we can get better results generated by segmentation CNN network.

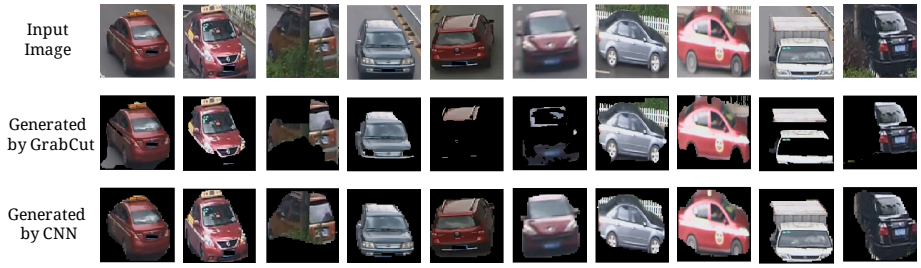

Fig. 1: **Qualitative results of the background-removed images.** The first row shows the input image and the second and third rows are the background-removed images generated by GrabCut and by the segmentation CNN network respectively.

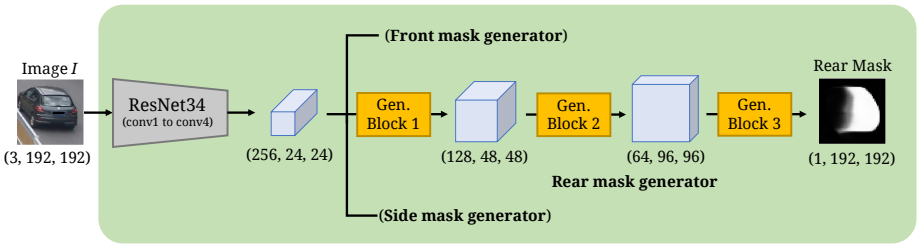

Fig. 2: **Model Architecture of our Semantic-guided Part Attention Network.**

2 Architecture of our Semantic-guided Part Attention Network

The network architecture of our proposed Semantic-guided Part Attention Network (SPAN) is shown in Fig. 2. It consists of a feature extractor which is the *conv*1 to *conv*4 in ResNet-34 [1] (CNN_{mask} in the main paper) and three mask generators (front, rear and side) with the same architecture, which only the rear mask generator is illustrated in details. Each generator contains three generative blocks (Gen. Block) and each block includes one transposed convolutional layer, batchnorm and ReLU layer. Considering that too powerful CNN model and extensive receptive field would lead to unexpected training results as described in the main paper, we only use the former four blocks in ResNet-34.

3 Selection of Hyper-parameters in Loss Functions.

We adopt three loss functions ($\mathcal{L}_{recon}$, $\mathcal{L}_{area}$ and $\mathcal{L}_{div}$) to supervise the training of our SPAN model. When computing the losses, there are two hyper-parameters should be selected, including max area ratio a in $\mathcal{L}_{area}$ and margin m in $\mathcal{L}_{div}$. The physical meaning and selection have been discussed in the main paper. To select

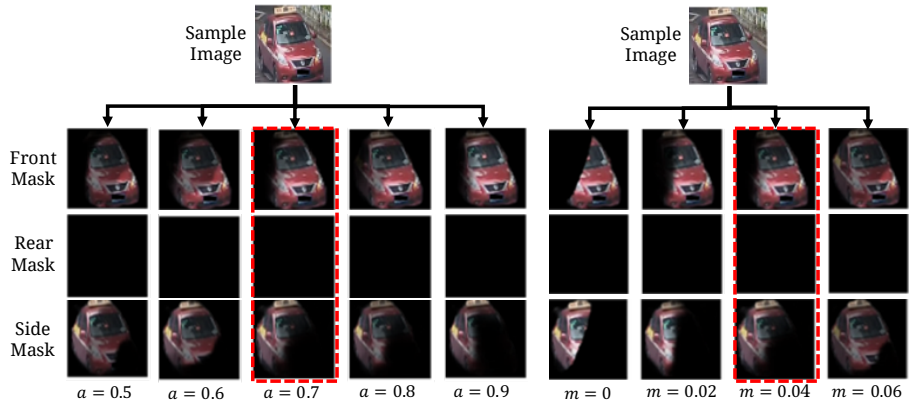

Fig. 3: **Analysis of the hyper-parameters: max area ratio a in $\mathcal{L}_{area}$ and margin m in $\mathcal{L}_{div}$.** We block the final selection of parameters in the red frames.

Table 1: **Parameters of $\mathcal{L}_{area}$.**

viewpoint	max area ratio a_l		
	front	rear	side
front	1	0	0
rear	0	1	0
side	0	0	1
front-side	0.7	0	0.7
rear-side	0	0.7	0.7

Table 2: **Parameters of $\mathcal{L}_{div}$.**

view pair	margin m
front, rear	0
front, side	0.04
rear, side	0.04

the hyper-parameters, we split a validation set out from the original training set of VeRi-776 dataset [2, 3] and observe the quality of generated attention masks of sampled images from validation set. We adjust one of the hyper-parameter while the other is fixed. The experiment results are shown in Fig 3.

The ideal part attention masks should cover all regional features which are belonging to their views while exclude the others. Take the results in Fig 3 as example, the front masks of $a = 0.8$ and $m = 0.06$ mistakenly include side views and the front mask of $m = 0.02$ incorrectly loses part of front view. Hence, based on the experiment results, we finally choose $a = 0.7$ for the visible views of two-view images and $m = 0.04$ for two adjacent views. The complete selection of hyper-parameters is shown in Table 1 and 2.

4 Other Qualitative Results of the Generated Part Masks

To verify the robustness of SPAN, we show more qualitative results of the generated part masks in Fig. 4. It is worth mentioning that the input images are all randomly chosen from the whole VeRi-776 dataset [2, 3] without manually selected.

Fig. 4: **Qualitative Results of Generated Part Masks.**

References

1. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778 (2016)
2. Liu, X., Liu, W., Ma, H., Fu, H.: Large-scale vehicle re-identification in urban surveillance videos. In: IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6 (2016)
3. Liu, X., Liu, W., Mei, T., Ma, H.: A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In: European Conference on Computer Vision (ECCV). pp. 869–884. Springer (2016)
4. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. ACM Transactions on Graphics (TOG) **23**(3), 309–314 (2004)