Visualizing the Invisible: Occluded Vehicle Segmentation and Recovery

Xiaosheng Yan[1],Feigege Wang[1],Wenxi Liu[1*],Yuanlong Yu[1*],Shengfeng He[2],Jia Pan[3]
[1] *College of Mathematics and Computer Science, Fuzhou University**
[2] *School of Computer Science and Engineering, South China University of Technology*
[3] *Department of Computer Science, The University of Hong Kong*

Abstract

In this paper, we propose a novel iterative multi-task framework to complete the segmentation mask of an occluded vehicle and recover the appearance of its invisible parts. In particular, firstly, to improve the quality of the segmentation completion, we present two coupled discriminators that introduce an auxiliary 3D model pool for sampling authentic silhouettes as adversarial samples. In addition, we propose a two-path structure with a shared network to enhance the appearance recovery capability. By iteratively performing the segmentation completion and the appearance recovery, the results will be progressively refined. To evaluate our method, we present a dataset, Occluded Vehicle dataset, containing synthetic and real-world occluded vehicle images. Based on this dataset, we conduct comparison experiments and demonstrate that our model outperforms the state-of-the-arts in both tasks of recovering segmentation mask and appearance for occluded vehicles. Moreover, we also demonstrate that our appearance recovery approach can benefit the occluded vehicle tracking in real-world videos.

1. Introduction

In recent years, segmentation techniques have made significant progress due to the development of deep learning [29, 49, 6, 16, 10, 5, 28]. Despite the achieved impressive performance, it is still difficult to accurately reason about objects under occlusions in an image. On the contrary, according to the study on amodal perception [21], one main strength of the human visual system is the ability to reason about the invisible, occluded parts of objects with high fidelity. To reduce the gap between the vision models and the human visual system, recent works start to investigate the problem of inferring the invisible part of objects, including amodal instance segmentation [25, 58] and generating the invisible part of objects [12].

In this paper, we focus on the task of appearance recovery for occluded vehicles. As we know, identifying vehicles is crucial for the applications of visual surveillance, intelligent traffic control, path prediction, and autonomous driving. However, in scenarios with vehicles and pedestrians, the occlusions are often observed and they increase the difficulty of learning visual representation of vehicles. As shown in Fig. 1, the tracker fails to follow the target under occlusions, since the occlusions prevent the tracker from learning the complete appearance representation of the target. In this situation, recovering the appearance of the invisible parts can mitigate such problem and benefit tracking.

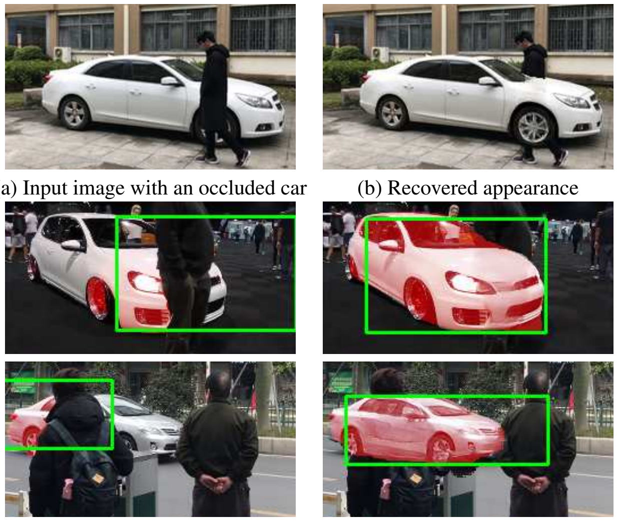
(a) Input image with an occluded car (b) Recovered appearance
(c) Tracking in original videos (d) Tracking in recovered videos

Figure 1: (a-b) Given the input image with an occluded car, our approach is capable of recovering the appearance of its invisible part. Illustration of tracking occluded vehicles in the original video (c) and the processed video (d) where the appearance of the target vehicles from occlusions have been recovered.

To accomplish the appearance recovery, we propose an iterative multi-task framework of segmentation completion and appearance recovery for occluded vehicles. Our framework consists of two modules: a segmentation completion network that aims at completing the incomplete segmentation mask of the occluded vehicle, and an appearance recovery network that aims to recover its appearance. In particular, to accurately recover the segmentation of the occluded vehicles, we propose two coupled discriminators, i.e., one object discriminator and one instance discriminator, in the segmentation completion network (see Fig. 2).

[*] Wenxi Liu and Yuanlong Yu are the corresponding authors.

The instance discriminator encourages the network to generate a segmentation mask similar to the ground-truth of the instance, while the object discriminator forces the produced mask to appear similar as a real vehicle. To accomplish this, we introduce an auxiliary 3D model pool to generate adversarial samples. Although the rendered 3D models are visually different from the real cars, their silhouettes are authentic compared with the real ones and thus they are suitable as the adversarial samples to further improve the generation quality. Since there are a large amount of vehicle models with varied types and poses in the 3D model pool, it implicitly brings richer prior into the segmentation completion.

In addition, to generate the visible parts from occlusions, we propose a two-path network architecture as the appearance recovery model (see Fig. 3). On the training stage, one path learns to fill in the colors of the invisible parts, while the other path is assigned with a more challenging task for inpainting the entire foreground vehicle given the image context. Since the parameters of networks on both paths are shared during training, the capability of the appearance recovery network will be enhanced. At test time, only the first path is deployed for generation. Lastly, our proposed multi-task framework allows the recovered image to be processed multiple times for refinement. To evaluate our method, we present an Occluded Vehicle dataset (OVD) that contains synthetic and real images. We test our approach on this dataset by comparing with the state-of-the-art methods in tasks of segmentation completion and appearance recovery, respectively. Moreover, we also apply our approach to recover the occluded vehicles in several real-world video sequences and demonstrate that our recovery approach can benefit the tracking as well. Hence, our contributions are summarized:

- We propose an iterative multi-task framework consisting of two separate modules for amodal segmentation of an occluded vehicle and appearance recovery.
- To infer the complete segmentation of a vehicle, we propose a segmentation completion network which has two coupled discriminators, which integrates an auxiliary 3D model pool to generate adversarial samples.
- We present a two-path appearance recovery network, which incorporates a foreground inpainting task with the appearance recovery of the invisible parts on the training stage to improve the recovery quality.
- We present a dataset, Occluded Vehicle Dataset (OVD), containing synthesized and real images of occluded vehicles for training and validation. Based on this dataset, we demonstrate that our work outperforms the state-of-the-art methods. Besides, we also collect several video sequences to demonstrate our approach can benefit occluded vehicle tracking.

2. Related Works

We survey the related literature on occlusion handling, generative adversarial network, and vehicle related works.

Occlusion handling. The occlusions are often observed in images or videos, which is often challenging in many vision problems. Thus, there are prior works studying the occlusion reasoning [48, 14, 43, 8, 18]. It has also been extensively studied in the detection and tracking community [23, 47, 40, 53, 19, 33], but these works do not consider recovering the appearance of the occluded objects. On the other hand, the amodal segmentation problem has been specifically presented and studied by [25, 58, 12, 13], which aim at providing a complete mask for occluded objects. Among the prior works, most similar to ours is Ehsani *et al.* [12], which presents a model, SeGAN, to generate the invisible parts of objects from indoor scenes. In their model, they deploy a residual network to produce a completed segmentation mask, which is too trivial to fully learn and recover the mask of the occluded objects with various shapes and poses. Besides, the resolution of their produced segmentation mask is much lower than that of the input image, which degrades its performance. Unlike SeGAN, we present an improved GAN model with two coupled discriminators to generate high-quality masks with the assistance of the silhouette masks sampled from various 3D models as adversarial samples.

Generative adversarial network. Generative adversarial network (GAN) is composed of a generative model and a discriminative model competing against each other in a two-player min-max game. It has been extensively studied [15, 7, 1] and widely applied in many applications, e.g. image-to-image translation [20, 57]. Besides, GAN has been applied in image inpainting [35, 50, 49]. SeGAN [12] also adopts GAN to generate the appearance of the invisible parts. However, their model requires the previously recovered segmentation mask as the only input, and thus it heavily relies on the quality of the input segmentation mask and lacks image contextual information. In our work, we present a two-path architecture integrated with an inpainting task that allows the network to thoroughly learn from the image context.

Vehicle related works. There are extensive studies on vehicles in the vision community, including detection [41, 46], tracking [32], counting [52], and re-identification [39, 45, 56]. In addition, with the rapid advancement in autonomous driving, more related research topics have been investigated [22, 2, 37, 31]. However, there are only a few recent works focusing on occluded vehicles, including vehicle detection under occlusion [54, 30, 4] and vehicle tracking with occlusions [34, 55]. Different from prior works, our work focuses on image-based occluded vehicle appearance recovery.

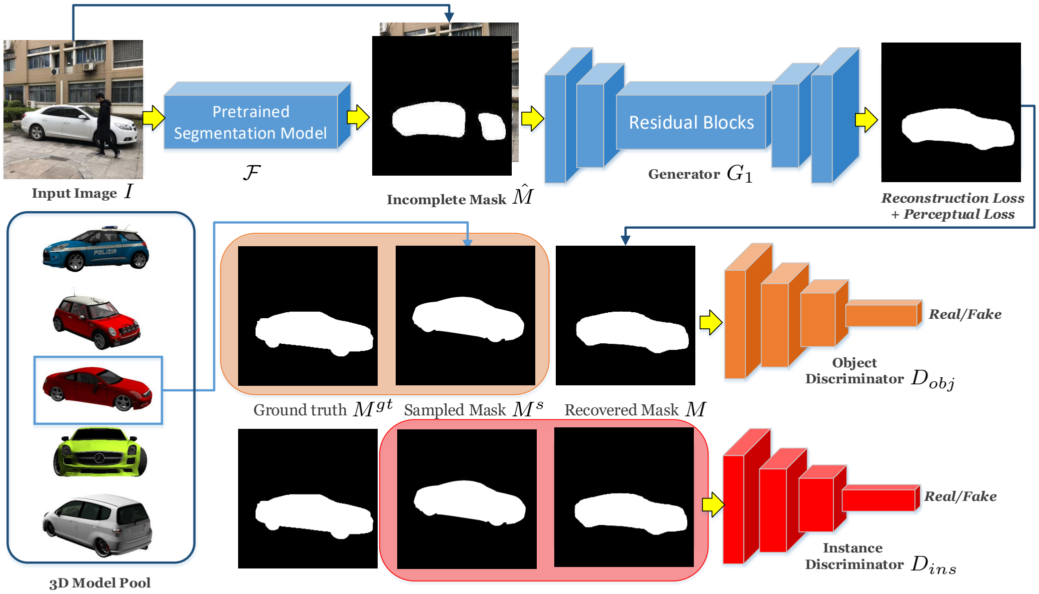
Input Image I
Pretrained Segmentation Model
$\mathcal{F}$
Incomplete Mask $\hat{M}$
Residual Blocks
Generator G_1
Reconstruction Loss + Perceptual Loss
Ground truth M^{gt}
Sampled Mask M^s
Recovered Mask M
Real/Fake
Object Discriminator D_{obj}
Real/Fake
Instance Discriminator D_{ins}
3D Model Pool

Figure 2: **Illustration of the segmentation completion network.** The input image I is passed to the pretrained segmentation model $\mathcal{F}$ and then concatenated with the computed incomplete mask $\hat{M}$ to produce the recovered segmentation mask M. In our framework, we present two coupled discriminators, both of which are fed with the same samples for different classification tasks. For the object discriminator D_{obj}, it aims to categorize the ground-truth M^{gt} and the sampled silhouette mask M^s as real and the recovered mask M as fake. For the instance discriminator D_{ins}, it aims to classify the ground-truth M^{gt} as real, while to classify the sampled silhouette mask M^s and the recovered mask M as fake. Specifically, the silhouette mask M^s is sampled from the 3D model pool as an adversarial sample.

3. Our Proposed Framework

3.1. Overview

Our framework is composed of two networks: the segmentation completion network (Fig. 2) and the appearance recovery network (Fig. 3). In particular, given the input image containing an occluded vehicle, the segmentation completion network generates the recovered segmentation mask. Then, the recovered segmentation mask is passed through the appearance recovery network to produce the invisible parts of vehicles. After painting the invisible parts back to the original image, the occluded vehicle will be on the foreground of the image (see Fig. 1(b)). Lastly, the image will be fed through the segmentation completion network and the appearance recovery network multiple times to refine and enhance the recovery quality. In the following subsections, we will introduce the details of both networks.

3.2. Segmentation Completion Network

As illustrated in Fig. 2, the segmentation completion network is based on a GAN model. In specific, the input image I containing occluded vehicles is first fed into a segmentation model $\mathcal{F}$ pretrained on public semantic segmentation datasets to generate the initial incomplete mask. The incomplete segmentation mask, $\hat{M}$, is then concatenated with the input image I and then passed into an encoder-decoder, i.e. the generator G_1. Its loss is the combination of the reconstruction loss (i.e. $\mathcal{L}_1$ loss) and the perceptual loss.

In a standard GAN, the generated sample and its corresponding ground-truth will be passed to a discriminator for classification, but the standard GAN tends to produce coarse results. To improve the quality of produced segmentation mask, we present two coupled discriminators including an object discriminator D_{obj} and an instance discriminator D_{ins}. In particular, the instance discriminator, same as the standard discriminator, is used to force the network to generate a mask identical with the ground truth. The object discriminator, on the other hand, aims to encourage the network to generate a mask similar to a real vehicle.

To accomplish this, we introduce an auxiliary 3D model pool, which collects a variety of rendered 3D vehicle models and their corresponding accurate silhouettes from ShapeNet [3]. Although there exists sim-to-real gap between the rendered 3D models and the real cars, the silhouettes of 3D models are visually similar to the real ones. And it is easy to extract accurate contour masks from the rendered images of these models. Thus, they can be used as the adversarial samples to improve the generation quality. By randomly selecting a silhouette as the adversarial sample, we collect three types of masks for discrimination, i.e. the ground-truth mask M^{gt}, the recovered mask M, and the sampled silhouette M^s. Inspired by the discriminator design of [51], as illustrated in Fig. 2, the object discrimi-

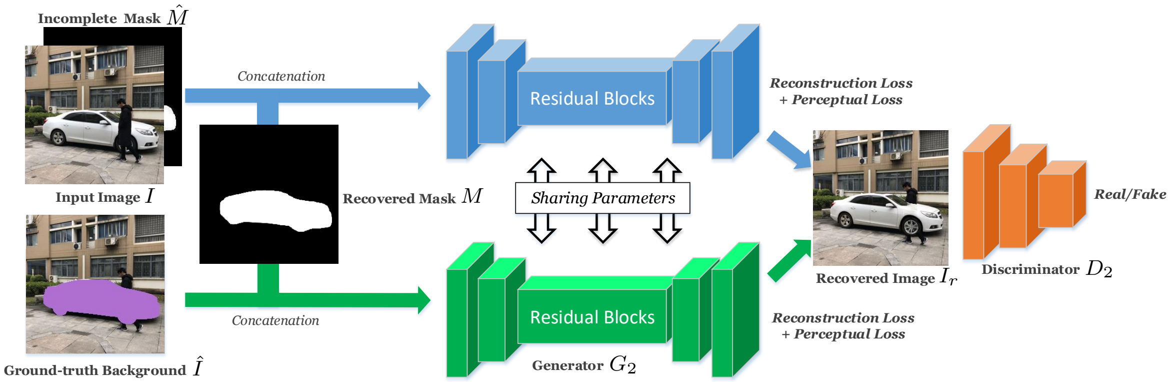

Figure 3: **Illustration of the appearance recovery network.** On the training stage, its generator has two paths to perform separate generation tasks but shares the same network. The first path aims at filling in colors for invisible parts. The second path aims at inpainting the image foreground given the image culling out the foreground object. At test time, only the first path is adopted for generating appearance.

nator aims to classify whether its input masks are real vehicle masks or not, i.e., to categorize the ground-truth and the sampled silhouette as real, and the recovered mask as fake, which is formulated as:

$$\mathcal{L}_{adv}(G_1, D_{obj}) = \mathbb{E}_{\hat{M}}[\log(1 - D_{obj}(G_1(I, \hat{M}))]+ \frac{1}{2}\left(\mathbb{E}_{M^{gt}}[\log D_{obj}(M^{gt})] + \mathbb{E}_{M^s}[\log D_{obj}(M^s)]\right), \quad (1)$$

The purpose of the instance discriminator is to classify whether the input mask is the segmentation of the vehicle, i.e., to categorize the ground-truth as real, but the sampled silhouette and the recovered mask as fake, which is defined:

$$\mathcal{L}_{adv}(G_1, D_{ins}) = \mathbb{E}_{M^{gt}}[\log D_{ins}(M^{gt}, I, \hat{M})]+ \frac{1}{2}(\mathbb{E}_{\hat{M}}[\log(1 - D_{ins}(G_1(I, \hat{M}), I, \hat{M}))]+ \mathbb{E}_{M^s}[\log(1 - D_{ins}(M^s, I, \hat{M}))]). \quad (2)$$

Note that D_{ins} uses the image I and mask $\hat{M}$ as additional inputs to attentively focus on the visible part of the occluded vehicle for discrimination. During training, a variety of different vehicles silhouettes are sampled, which encourages the network to learn the representative feature of real vehicle masks (e.g. the positions and shapes of wheels), and thus to encourage the produced masks to appear similar as real vehicles. Hence, the final objective of the segmentation completion network is to minimize:

$$\mathcal{L}^{seg} = \mathcal{L}_{adv}(G_1, D_{obj}) + \mathcal{L}_{adv}(G_1, D_{ins})+ \lambda\mathcal{L}_{L1}(G_1) + \beta\mathcal{L}_{perc}(G_1), \quad (3)$$

where $\mathcal{L}_{L1}(\cdot)$ and $\mathcal{L}_{perc}(\cdot)$ denote the reconstruction loss and the perceptual loss, respectively.

3.3. Appearance Recovery Network

With the recovered segmentation mask M, our framework aims to generate the appearance of the invisible parts for the occluded vehicle in the next step. As illustrated in Fig. 3, the appearance recovery network is also based on a GAN model.

As the generator, we propose a two-path network architecture which performs two separate tasks while sharing the same network G_2. For the first path, the recovered segmentation mask M is concatenated with the input image I and the incomplete mask $\hat{M}$, to fed into the network. Since the incomplete mask indicates the visible parts and the recovered mask estimates the silhouette of the whole unoccluded vehicle, the purpose of this path is to fill in the colors for invisible parts.

In addition, the second path is supervised to perform a more challenging task, i.e., to inpaint the whole vehicle based on image context. It receives the concatenation of the recovered mask M and the ground-truth background $\hat{I}$ along with a zero map ϕ for padding in order to learn for inpainting. When training, the network G_2 is shared on both paths, so it will be endowed with the ability that not only recovers the invisible parts but also the entire vehicle based on contextual information, which significantly improves the capability of the generator. The recovered images are sent to the discriminator D_2 to guarantee the image quality. Thus, the objective of the appearance recovery network is:

$$\mathcal{L}^{app} = \mathcal{L}_{adv}(G_2(I, \hat{M}, M), D_2) + \mathcal{L}_{adv}(G_2(\hat{I}, \phi, M), D_2)+ \lambda_1\mathcal{L}_{L1}(G_2(I, \hat{M}, M)) + \beta_1\mathcal{L}_{perc}(G_2(I, \hat{M}, M))+ \lambda_2\mathcal{L}_{L1}(G_2(\hat{I}, \phi, M)) + \beta_2\mathcal{L}_{perc}(G_2(\hat{I}, \phi, M)).$$

On the testing stage, since the ground-truth background of the test image is unknown, the second path is disabled and the generator on the first path is applied only.

3.4. Iterative Refinement

On the testing stage, the recovered image can be produced by passing the input image I through both generators, i.e., $I_r = G_2(G_1(\mathcal{F}(I)))$, where $\mathcal{F}$ refers to the pretrained segmentation model. With the invisible parts of the vehicle recovered in the input image, the occluded vehicle in the image will appear on the foreground. However, there may exist artifacts in the recovered image I_r. Our multi-task framework allows the recovered image to be processed multiple times and finally produces a refined image. For example, the synthesized image in the second iteration is produced as: $I_r^{(2)} = G_2(G_1(\mathcal{F}(I_r^{(1)})))$, where $I_r^{(1)} = I_r$. The intuition of the recursive process is based on the correlation between the segmentation completion and the appearance recovery. In each iteration, the completeness of the recovered segmentation mask affects the quality of the appearance recovery. In the same iteration, the appearance recovery network recovers the appearance while implicitly refining its segmentation mask. Hence, the iterative process may progressively improve the quality of generation, as shown by the example in Fig. 6.

4. Experiments

4.1. Implementation details

3D model pool. From ShapeNet [3], we select 401 different classes of vehicles and, for each vehicle, we screenshot each rendered image from 80 different viewpoints. Since the background of the rendered image is very clean, we can simply extract the accurate silhouettes by thresholding. In this way, we collect 32,080 silhouettes to form the auxiliary 3D model pool.

Network structure and training. In practice, as encoder-decoder structure, both G_1 and G_2 downsample the resolution from 256 to 64 and then upsample to the original spatial resolution. As the middle layers, there are 8 residual blocks with the dilation rate 2. D_2 adopts the Patch-GAN discriminator [20], while D_{obj} and D_{ins} use the same structure except that they have an additional fully connected layer for classification. For hyper-parameters, we set $\lambda = \lambda_1 = \lambda_2 = 10$ and $\beta = \beta_1 = \beta_2 = 1$. In practice, we employ [16] as the pretrained segmentation model $\mathcal{F}$. Our model is implemented in Tensorflow on PC with Intel Core i7-6700 CPU, 32GB RAM, and a single NVIDIA Titan Xp. We first train the segmentation completion network and the appearance recovery network separately using 256×256 images with a batch size of 4, with Adam solver. To train each network, we set the learning rate as 10^{-4} until the loss plateaus and then lower it to 10^{-5} until convergence. We then train both networks in an end-to-end manner with the learning rate 10^{-6}.

Metrics. We evaluate our methods in two tasks. For the recovered segmentation mask, we adopt precision, recall, F1-score, Intersection over Union (IoU), the per-pixel $\mathcal{L}_1$ error, and the per-pixel $\mathcal{L}_2$ error as the evaluation metrics. For the recovered appearance of the vehicle, we adopt the per-pixel $\mathcal{L}_1$ error and per-pixel $\mathcal{L}_2$ error. Additionally, to evaluate the generation quality, the inception score [38] is often used. Since we care about the generation quality rather than the diversity, we simply adopt the conditional probability computed by the pretrained Inception [42] for evaluating the recovered vehicles, denoted as *Inception conditional probability* (ICP). Besides, similar to the inception score and the FCN-score used in [20], we also apply the state-of-the-art segmentation model [16] trained on Cityscape [9] to segment the whole recovered image with the reference of the ground-truth labels. The intuition is that, if the recovered vehicle in the image is realistic, the segmentation model trained on real images will be able to classify it correctly. Thus, its segmentation accuracy for vehicles is adopted as another metric, denoted as *Segmentation Score* (SS).

4.2. Occluded Vehicle Dataset

To our best knowledge, there is no public dataset providing occluded vehicles with unoccluded segmentation and appearance ground-truth. For experiments, we present a new dataset called Occluded Vehicle Dataset (OVD). To produce synthesized data, we leverage the images of the Cars dataset [24] as the base images for synthesis. The original dataset contains 16,185 images with 196 classes of cars. Since most cars from the dataset are unoccluded, we manually label the segmentation of the cars as ground truth. Besides, we randomly place some real pedestrians, vehicles, and other objects, which are randomly cropped from COCO [26] and CityScape [9], as occlusions over the vehicles of these base images. Then, we adopt the Deep Harmonization technique [44] to make these synthetic images look natural. In this way, we collect a total of 33,100 images for training and 1000 images for testing. Note that the vehicles in the test images are unseen in the training set. In addition, we also collect and label 100 real-world images as part of the dataset for testing. Therefore, we enrich the diversity of the images in four aspects: (1) the number of vehicles; (2) the classes of vehicles; (3) the poses of vehicles; (4) the types of the occlusions. Moreover, we collect and label 4 video sequences (i.e. Vid-1, Vid-2, Vid-3, Vid-4) with occluded vehicles, which are captured from underground parking lots, crowded exhibitions, and streets. More examples of OVD are shown in the experiment section.

4.3. Ablation studies

To analyze our proposed framework, we first evaluate our proposed discriminators in the segmentation completion network, and then evaluate the two-path structure of the appearance recovery network and the effect of recovered mask for appearance recovery. Lastly, we analyze our iterative generation method. All the ablation studies are performed on the synthetic and real images of the testset in OVD.

Structure	Input	L1 ↓	L2 ↓	F1 ↑	IoU ↑
$D_{standard}$ $\{D_{obj}, D_{ins}\}$	Syn.	0.0702 **0.0559**	0.0667 **0.0535**	0.8403 **0.8798**	0.7421 **0.7939**
$D_{standard}$ $\{D_{obj}, D_{ins}\}$	Real	0.0338 **0.0322**	0.0335 **0.0314**	0.8890 **0.8981**	0.8067 **0.8193**
Structure	Input	L1 ↓	L2 ↓	ICP ↑	SS ↑
one-path two-path	Syn.	0.0421 **0.0364**	0.0181 **0.0161**	0.6214 **0.6676**	0.8350 **0.9411**
one-path two-path	Real	0.0201 **0.0171**	0.0077 **0.0063**	0.8058 **0.8216**	0.9131 **0.9292**
Structure	Input	L1 ↓	L2 ↓	ICP ↑	SS ↑
w/o Mask w/ Mask	Real	0.0325 **0.0173**	0.0124 **0.0063**	0.6716 **0.8350**	0.8849 **0.9356**

Table 1: Ablation experiments for our architectures.

		Segmentation recovery				Appearance recovery			
Iter.	Input	L1 ↓	L2 ↓	F1 ↑	IoU ↑	L1 ↓	L2 ↓	ICP ↑	SS ↑
1	Syn.	0.0559	0.0535	0.8798	0.7939	0.0364	0.0161	0.6676	0.9411
2		**0.0499**	**0.0480**	**0.8935**	**0.8137**	**0.0341**	**0.0146**	**0.6765**	**0.9545**
3		0.0510	0.0493	0.8902	0.8080	0.0343	0.0148	0.6748	0.8458
1	Real	0.0322	**0.0314**	0.8890	0.8066	**0.0171**	**0.0063**	0.8216	0.9292
2		**0.0320**	**0.0314**	**0.8898**	**0.8067**	0.0173	**0.0063**	**0.8350**	**0.9356**
3		0.0342	0.0336	0.8810	0.7926	0.0178	0.0066	0.8206	0.9314

Table 2: Ablation experiments for studying the iterations of our model in the tasks of segmentation and appearance recovery.

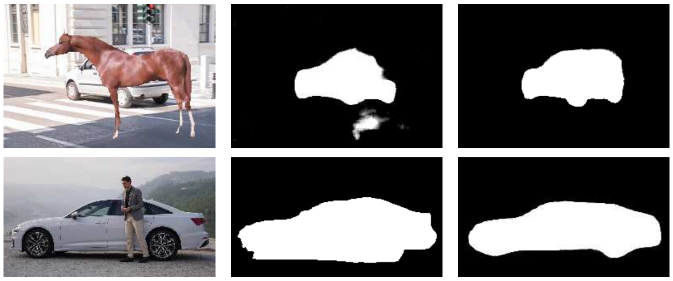
Input image $D_{standard}$ $\{D_{obj}, D_{ins}\}$

Figure 4: Generated complete segmentation masks on exemplar synthetic and real images for evaluating our discriminators.

By evaluating the discriminators in the segmentation completion network, we compare our proposed model ($\{D_{obj}, D_{ins}\}$) against the standard discriminator $D_{standard}$ that is a single discriminator network for real/fake classification. As illustrated in Tab. 1, the quality of segmentation completion from our proposed model is generally improved. As shown in Fig. 4, the completed segmentation masks may be coarse and noisy using $D_{standard}$.

To demonstrate the effectiveness of the two-path structure, we compare our two-path structure with the one-path structure which contains the first path only. The second path requires the ground-truth labels, so it cannot be applied in test solely. In Tab. 1, the two-path structure shows the obvious advantages over its counterpart, as one-path may not be capable enough to fully recover the appearance from the invisible parts, as shown in Fig.5. Furthermore, Tab. 1 depicts the recovered masks that strenghten appearance recovery.

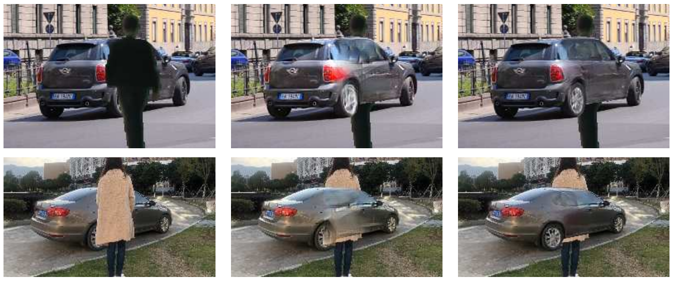
Input image One-path structure Two-path structure

Figure 5: Recovered appearance on exemplar synthetic and real images for evaluating our proposed two-path structure.

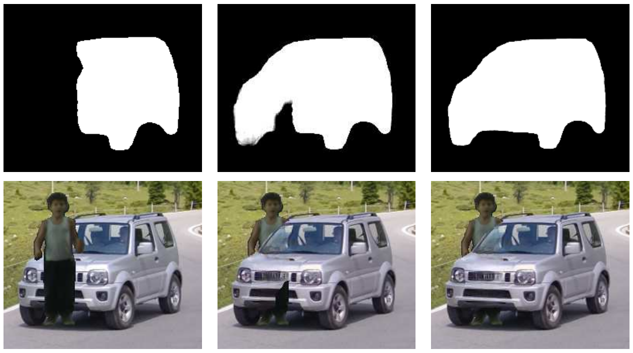
Input 1^{st} iteration 2^{nd} iteration

Figure 6: Illustration of an example on our iterative refinement. The first column refers the input image and its corresponding incomplete mask. The second and third column refer to the results produced at the first and the second iteration, respectively.

Lastly, we analyze the performance of running our model for 1, 2, and 3 iterations in the tasks of segmentation completion and appearance recovery. In specific, the model running for 1 iteration refers to the process of passing input images through two generators once. Generally, we obtain the optimal performance in the second iteration. For synthetic images, due to different kinds of synthetic occlusions in images, our model requires multiple iterations to progressively remove the occlusions and recover the missing details. We show an example of the progressive refinement in Fig. 6. For real images with less severe occlusions, the second iteration only slightly improves the performance of recovering segmentation and appearance, since the model on the first iteration has already produced recognizable shapes. But its improvement on ICP indicates that the iterative process still manages to refine the appearance for the recovered object. Besides, based on our observation, the performance will degrade for more than 3 iterations. This is because, as the results cannot be further refined, the error of both stages will be accumulated with more iterations.

	Synthetic images						*Real images*						*Real vehicles in* [26]			
Model	Prec. ↑	Recall ↑	F1 ↑	IoU ↑	L1 ↓	L2 ↓	Prec. ↑	Recall ↑	F1 ↑	IoU ↑	L1 ↓	L2 ↓	F1 ↑	IoU ↑	L1 ↓	L2 ↓
Mask R-CNN [16]	0.7803	0.7914	0.7653	0.6246	0.1174	0.1162	0.7066	0.9197	0.7953	0.6619	0.0731	0.0730	0.7577	0.6213	0.1281	0.1281
Deeplab [6]	0.8810	0.8779	0.8794	0.7937	0.0574	0.0559	0.8918	0.7508	0.8111	0.6976	0.0545	0.0544	0.7459	0.6213	0.1182	0.1182
SharpMask [36]	0.8286	**0.9404**	0.8751	0.7840	0.0646	0.0635	0.8025	**0.9518**	0.8669	0.7693	0.0463	0.0462	0.7903	0.6627	0.1509	0.1266
pix2pix [20]	0.8865	0.8906	**0.8821**	0.7932	0.0575	0.0557	0.8471	0.9055	0.8718	0.7763	0.0414	**0.0407**	0.7726	0.6429	0.1224	0.1218
SeGAN [12]	0.7931	**0.9016**	0.8367	0.7236	0.0846	0.0835	0.7477	**0.9417**	0.8303	0.7133	0.0603	0.0602	0.8085	0.6894	0.1123	0.1123
Ours (1^{st} iter.)	**0.9590**	0.8229	0.8798	**0.7939**	**0.0559**	**0.0535**	**0.9821**	0.8176	**0.8890**	**0.8067**	**0.0322**	**0.0314**	**0.8190**	**0.7113**	**0.0921**	**0.0904**
Ours (2^{nd} iter.)	**0.9625**	0.8416	**0.8935**	**0.8137**	**0.0499**	**0.0480**	**0.9854**	0.8148	**0.8898**	**0.8066**	**0.0320**	**0.0314**	**0.8234**	**0.7133**	**0.0915**	**0.0893**

Table 3: The comparison results of segmentation completion in Occluded Vehicle dataset. On each column, the top performer is marked in red while the second one is marked in blue.

4.4. Results analysis

We compare our model with the state-of-the-art methods in two tasks, i.e. segmentation completion and appearance recovery. For the task of segmentation completion, we compare with Mask R-CNN [16], Deeplab [6], SharpMask [36], pix2pix [20], and SeGAN [12]. For appearance recovery, we compare with Deepfill [50], Liu *et al.*[27], Pathak *et al.* [35], pix2pix, and SeGAN. Specifically, Mask R-CNN and Deeplab are the state-of-the-art segmentation models. SharpMask has been proposed to complete and refine the generated masks. As a supervised GAN model, pix2pix can be applied in several applications, including segmentation and image synthesis. Deepfill [50], [27], [35] are the state-of-the-art inpainting methods and SeGAN claims to achieve the state-of-the-art performance in both amodal segmentation and appearace recovery. In experiments, deeplab, pix2pix, and Mask R-CNN are trained from scratch, while SharpMask, SeGAN, and the inpainting models are finetuned on the pre-trained models that gains better performance. Besides, we run our model for 1 iteration and 2 iterations respectively for comparison. The evaluations are separately performed on the synthetic images and the real images of our dataset.

Segmentation completion. Aside from OVD, we also select the real vehicles from [26] for evaluation. We demonstrate the comparison results in Tab. 3. Generally, the result shows that our model with or without the iterative refinement outperforms the prior methods. As illustrated in Fig. 7, our model can produce masks with smooth contours and clear shapes of wheels and bodywork, due to the involvement of the object discriminator and the auxiliary 3D model pool. The results of Deeplab are comparable to ours, but the shapes of wheels and bodyworks are not clear. Since SeGAN generates masks with the low-resolution (i.e. 58 × 58) and upsamples them to 256 × 256, their results appear to be coarse. SharpMask and pix2pix can produce more complete and finer masks than SeGAN, but their contours are not smooth enough.

Model	Type	Input	L1 ↓	L2 ↓	ICP ↑	SS ↑
Deepfill [50]	Syn.	M^{gt}	0.0284	0.0107	0.5620	0.8295
Liu *et al.* [27]			0.0272	0.0074	0.6284	0.8672
Pathak *et al.* [35]			0.0207	0.0088	0.5708	0.8517
pix2pix [20]			0.0174	0.0060	0.7081	0.9410
SeGAN [12]			0.0181	0.0055	0.6662	0.9371
Ours (1^{st} iter.)			**0.0159**	**0.0038**	**0.7436**	**0.9458**
Ours (2^{nd} iter.)			**0.0158**	**0.0039**	**0.7267**	**0.9447**
pix2pix [20]	Syn.	M	0.0455	0.0226	0.6337	0.8825
SeGAN [12]			0.0499	0.0224	0.6138	0.9165
Ours (1^{st} iter.)			**0.0364**	**0.0161**	**0.6676**	**0.9411**
Ours (2^{nd} iter.)			**0.0341**	**0.0146**	**0.6765**	**0.9545**
pix2pix [20]	Real	M	0.0182	**0.0074**	0.7888	0.9165
SeGAN [12]			0.0256	0.0114	0.4984	0.9192
Ours (1^{st} iter.)			**0.0171**	**0.0063**	**0.8216**	**0.9292**
Ours (2^{nd} iter.)			**0.0173**	**0.0063**	**0.8350**	**0.9356**

Table 4: The comparison results of Appearance recovery for the synthetic and real images. In order to perform fair comparisons, we assign the ground-truth segmentation mask M^{gt} and the predicted recovered segmentation mask M as the inputs, respectively.

Appearance recovery. The comparison results are shown in Tab. 4. Since Deepfill requires the image context without vehicles, we provide the ground-truth segmentation mask M^{gt}. For fair comparison, we also provide the ground-truth masks for the other methods as well. As observed, our method demonstrates superior performance over others. However, since the ground-truth masks are provided, the iterative refinement does not show much effect and even degrades the model performance a little. In addition, we perform comparisons of appearance recovery for methods given their predicted masks. These comparisons are performed in both synthetic and real images. According to ICP for evaluating the recovered vehicles and SS for evaluating the recovered image, our method generates more plausible images. As shown in Fig. 8, only Deepfill is provided with the ground-truth mask, so its recovered vehicles have more complete shape but without much details. The other comparison results are generated based on the input images or their predicted masks. Due to our appearance recovery network, we can paint more details on the invisible parts. Taking the first and the third row in Fig. 8 as examples, our model manages to paint a wheel on the proper position in the image. More recovered examples from [11, 58] and OVD are illustrated in Fig. 9.

Occluded vehicle tracking. We apply our method in four real-world videos (Vid-1, Vid-2, Vid-3, Vid-4) to recover the vehicles to be unoccluded. Then, we apply the same tracker KCF [17] to track the vehicles from the original

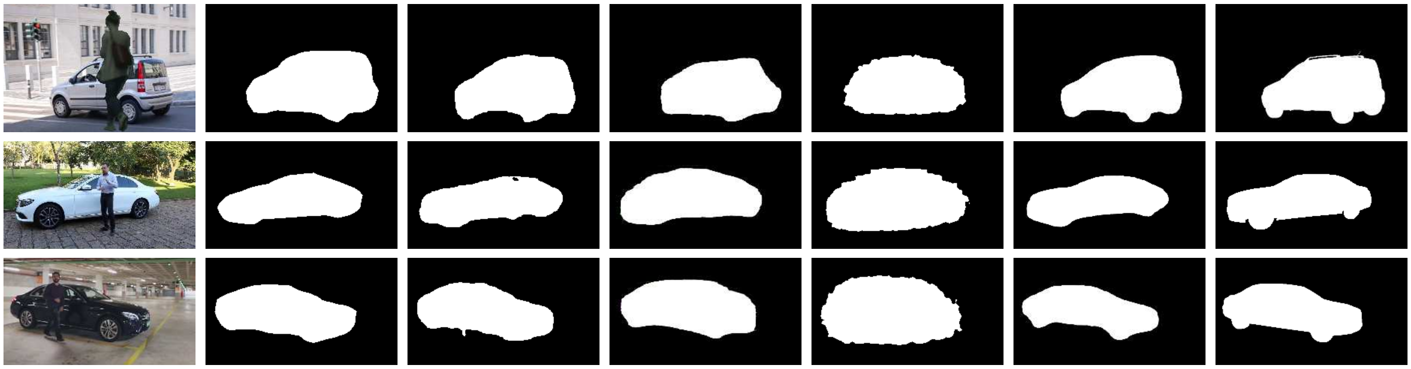

Figure 7: Examples of the segmentation completion comparison.

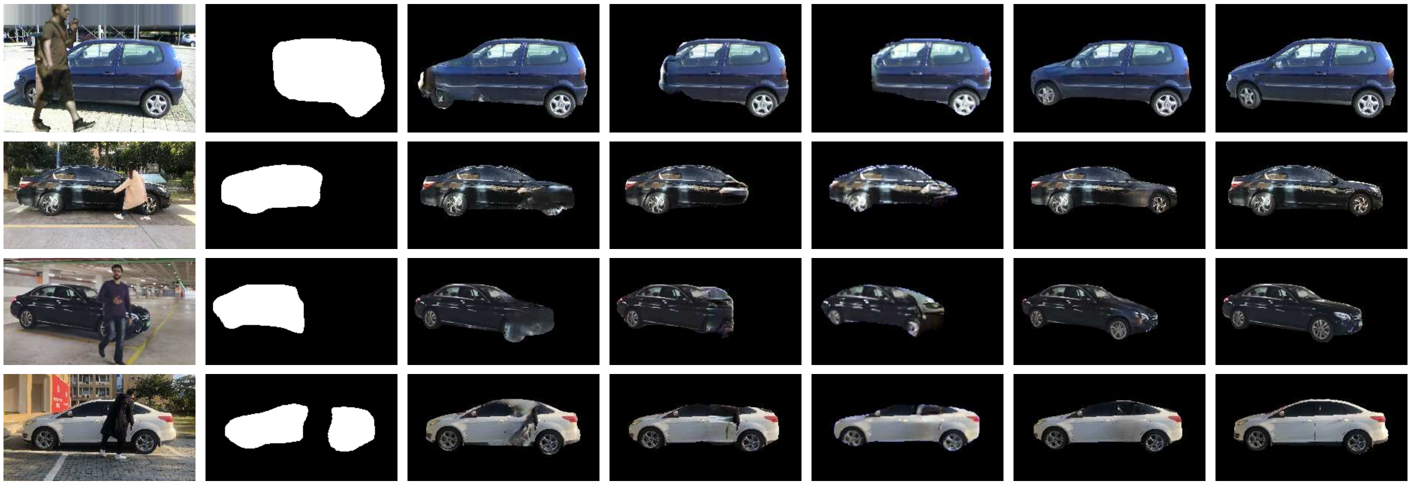

Figure 8: Examples of the appearance recovery comparison.

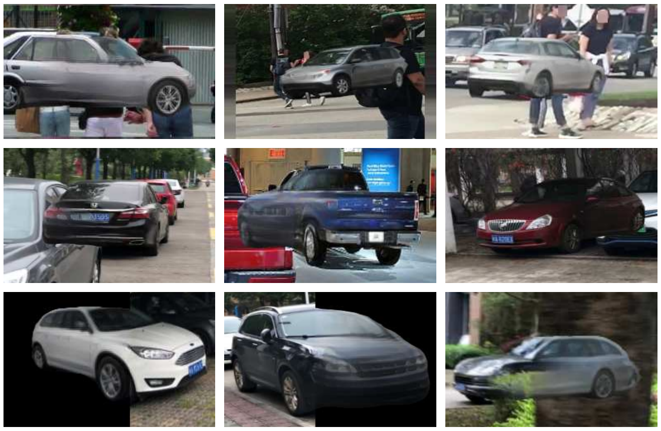

Figure 9: Examples of recovered real occluded vehicles from public datasets [11, 58] and our OVD dataset.

videos and the recovered videos, respectively. The results are illustrated in Tab. 5 in terms of average pixel error (APE) and average overlap (AO) which indicates that our recovered videos benefit the vehicle tracking (see Fig. 1).

	APE ↓		AO ↑	
	Original	Recovered	Original	Recovered
Vid-1	34.60	**8.32**	0.6489	**0.8072**
Vid-2	26.30	**15.83**	0.7285	**0.8040**
Vid-3	87.71	**21.51**	0.3584	**0.6755**
Vid-4	7.29	**5.67**	0.7494	**0.7497**

Table 5: Tracking performance comparison for the original and the recovered videos. The better number in comparison is in bold.

4.5. Conclusion

In this paper, we propose an iterative multi-task framework to recover the segmentation mask and the appearance for occluded vehicles. In particular, we propose two coupled discriminators and a two-path structure with a shared network and evaluate our method in a proposed dataset. Moreover, we show our method can benefit the occluded vehicle tracking.

Acknowledgement. Our project is supported by the National Natural Science Foundation of China (No. 61702104, 61873067, and 61702194), the Natural Science Foundation of Fujian Province (No. 2016J05155), and the Guangzhou Key Industrial Technology Research fund (No. 201802010036).

References

[1] Martn Arjovsky, Soumith Chintala, and Lon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 2

[2] Aseem Behl, Omid Hosseini Jafari, Siva Karthik Mustikovela, Hassan Abu Alhaija, Carsten Rother, and Andreas Geiger. Bounding boxes, segmentations and object coordinates: How important is recognition for 3d scene flow estimation in autonomous driving scenarios? In *ICCV*, 2017. 2

[3] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], 2015. 3, 5

[4] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan. Vision-based occlusion handling and vehicle classification for traffic surveillance systems. *IEEE Intelligent Transportation Systems Magazine*, 2018. 2

[5] Liang-Chieh Chen, Alexander Hermans, George Papandreou, Florian Schroff, Peng Wang, and Hartwig Adam. Masklab: Instance segmentation by refining object detection with semantic and direction features. In *CVPR*, 2018. 1

[6] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *TPAMI*, 2018. 1, 7, 8

[7] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: interpretable representation learning by information maximizing generative adversarial nets. *NeurIPS*, 2016. 2

[8] Yi-Ting Chen, Xiaokai Liu, and Ming-Hsuan Yang. Multi-instance object segmentation with occlusion handling. In *CVPR*, 2015. 2

[9] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 5

[10] Jifeng Dai, Kaiming He, and Jian Sun. Instance-aware semantic segmentation via multi-task network cascades. In *CVPR*, 2016. 1

[11] N. Dinesh Reddy, Minh Vo, and Srinivasa G. Narasimhan. Carfusion: Combining point tracking and part detection for dynamic 3d reconstruction of vehicles. In *CVPR*, 2018. 7, 8

[12] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *CVPR*, 2018. 1, 2, 7, 8

[13] Patrick Follmann, Rebecca Konig, Philipp Hartinger, and Michael Klostermann. Learning to see the invisible: End-to-end trainable amodal instance segmentation. *arXiv preprint arXiv:1804.08864*, 2018. 2

[14] Tianshi Gao, Benjamin Packer, and Daphne Koller. A segmentation-aware object detection model with occlusion handling. In *CVPR*, 2011. 2

[15] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 2

[16] Kaiming He, Georgia Gkioxari, Piotr Dollr, and Ross B. Girshick. Mask r-cnn. In *ICCV*, 2017. 1, 5, 7

[17] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 2015. 7

[18] Edward Hsiao and Martial Hebert. Occlusion reasoning for object detection under arbitrary viewpoint. In *CVPR*, 2012. 2

[19] Yang Hua, Karteek Alahari, and Cordelia Schmid. Occlusion and motion reasoning for long-term tracking. In *ECCV*, 2014. 2

[20] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2, 5, 7, 8

[21] Gaetano Kanizsa, Paolo Legrenzi, and Paolo Bozzi. Organization in vision : essays on gestalt perception. 1979. 1

[22] Jinkyu Kim and John F. Canny. Interpretable learning for self-driving cars by visualizing causal attention. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017. 2

[23] Dieter Koller, Joseph Weber, and Jitendra Malik. Robust multiple car tracking with occlusion reasoning. In *ECCV*, 1994. 2

[24] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International IEEE Workshop on 3D Representation and Recognition*, Sydney, Australia, 2013. 5

[25] Ke Li and Jitendra Malik. Amodal instance segmentation. *ECCV*, 2016. 1, 2

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollr, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ECCV*, 2014. 5, 7

[27] Guilin Liu, Fitsum A. Reda, Kevin J. Shih, Ting Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. 2018. 7

[28] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. In *CVPR*, 2018. 1

[29] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

[30] Xiubo Ma and Xiongwei Sun. Detection and segmentation of occluded vehicles based on symmetry analysis. In *International Conference on Systems and Informatics*, 2017. 2

[31] Ana I. Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso N. Garca, and Davide Scaramuzza. Event-based vision meets deep learning on steering prediction for self-driving cars. In *CVPR*, 2018. 2

[32] Xue Mei and Haibin Ling. Robust visual tracking and vehicle classification via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2011. 2

[33] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCVW*, 2017. 2

[34] Clement Chun Cheong Pang, William Wai Leung Lam, and Nelson Hon Ching Yung. A novel method for resolving vehicle occlusion in a monocular traffic-image sequence. *IEEE Transactions on Intelligent Transportation Systems*, 2004. 2

[35] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016. 2, 7

[36] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *ECCV*, 2016. 7, 8

[37] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *CVPR*, 2018. 2

[38] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training GANs. In *NeurIPS*, 2016. 5

[39] Yantao Shen, Tong Xiao, Hongsheng Li, Shuai Yi, and Xiaogang Wang. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In *ICCV*, 2017. 2

[40] Guang Shu, Afshin Dehghan, Omar Oreifej, Emily Hand, and Mubarak Shah. Part-based multiple-person tracking with partial occlusion handling. In *CVPR*, 2012. 2

[41] Sayanan Sivaraman and Mohan Manubhai Trivedi. Looking at vehicles on the road: A survey of vision-based vehicle detection, tracking, and behavior analysis. *IEEE Transactions on Intelligent Transportation Systems*, 2013. 2

[42] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 5

[43] Joseph Tighe, Marc Niethammer, and Svetlana Lazebnik. Scene parsing with object instances and occlusion ordering. In *CVPR*, 2014. 2

[44] Y.-H. Tsai, X. Shen, Z. Lin, K. Sunkavalli, X. Lu, and M.-H. Yang. Deep image harmonization. In *CVPR*, 2017. 5

[45] Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao, Junjie Yan, Shengjin Wang, Hongsheng Li, and Xiaogang Wang. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In *ICCV*, 2017. 2

[46] Ke Yan, Yonghong Tian, Yaowei Wang, Wei Zeng, and Tiejun Huang. Exploiting multi-grain ranking constraints for precisely searching visually-similar vehicles. In *ICCV*, 2017. 2

[47] Tao Yang, Quan Pan, Jing Li, and S.Z. Li. Real-time multiple objects tracking with occlusion handling in dynamic scenes. In *CVPR*, 2005. 2

[48] Yi Yang, Sam Hallman, Deva Ramanan, and Charless C. Fowlkes. Layered object models for image segmentation. *TPAMI*, 2012. 2

[49] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *ICLR*, 2016. 1, 2

[50] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018. 2, 7, 8

[51] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. Stack-GAN++: Realistic image synthesis with stacked generative adversarial networks. *TPAMI*, 2018. 3

[52] Shanghang Zhang, Guanhang Wu, Joo P. Costeira, and Jos M. F. Moura. Fcn-rlstm: Deep spatio-temporal neural networks for vehicle counting in city cameras. In *ICCV*, 2017. 2

[53] Tianzhu Zhang, Kui Jia, Changsheng Xu, Yi Ma, and Narendra Ahuja. Partial occlusion handling for visual tracking via robust part matching. In *CVPR*, 2014. 2

[54] Weigang Zhang, Q.m.j. Wu, Xiaokang Yang, and Xiangzhong Fang. Multilevel framework to detect and handle vehicle occlusion. *IEEE Transactions on Intelligent Transportation Systems*, 2008. 2

[55] Na Zhao, Yingjie Xia, Chao Xu, Xingmin Shi, and Yuncai Liu. Appos: An adaptive partial occlusion segmentation method for multiple vehicles tracking. *Journal of Visual Communication and Image Representation*, 2016. 2

[56] Yi Zhouy and Ling Shao. Viewpoint-aware attentive multiview inference for vehicle re-identification. In *CVPR*, 2018. 2

[57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2

[58] Yan Zhu, Yuandong Tian, Dimitris N. Metaxas, and Piotr Dollr. Semantic amodal segmentation. In *CVPR*, 2017. 1, 2, 7, 8