Human De-occlusion: Invisible Perception and Recovery for Humans

Qiang Zhou[1]*, Shiyin Wang[2], Yitong Wang[2], Zilong Huang[1], Xinggang Wang[1]†
[1]School of EIC, Huazhong University of Science and Technology [2]ByteDance Inc.
theodoruszq@gmail.com shiyinwang.ai@bytedance.com wangyitong@pku.edu.cn
zilong.huang2020@gmail.com xgwang@hust.edu.cn

Abstract

In this paper, we tackle the problem of human de-occlusion which reasons about occluded segmentation masks and invisible appearance content of humans. In particular, a two-stage framework is proposed to estimate the invisible portions and recover the content inside. For the stage of mask completion, a stacked network structure is devised to refine inaccurate masks from a general instance segmentation model and predict integrated masks simultaneously. Additionally, the guidance from human parsing and typical pose masks are leveraged to bring prior information. For the stage of content recovery, a novel parsing guided attention module is applied to isolate body parts and capture context information across multiple scales. Besides, an Amodal Human Perception dataset (AHP) is collected to settle the task of human de-occlusion. AHP has advantages of providing annotations from real-world scenes and the number of humans is comparatively larger than other amodal perception datasets. Based on this dataset, experiments demonstrate that our method performs over the state-of-the-art techniques in both tasks of mask completion and content recovery. Our AHP dataset is available at https://sydney0zq.github.io/ahp/.

1. Introduction

Visual recognition tasks have witnessed significant advances driven by deep learning, such as classification [21, 14], detection [38, 10] and segmentation [28, 51]. Despite the achieved progress, *amodal perception*, *i.e.* to reason about the occluded parts of objects, is still challenging for vision models. In contrast, it is easy for humans to interpolate the occluded portions with human visual systems [59, 34]. To reduce the recognition ability gap between the models and humans, recent works have proposed methods to infer the occluded parts of objects, including estimating the invisible segmentation [25, 59, 37, 50, 54] and recovering the invisible content of objects [6, 50, 54].

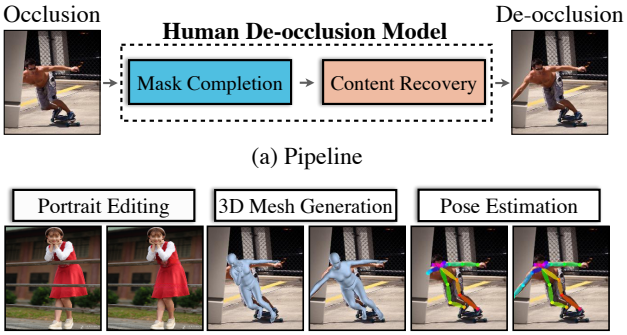

Figure 1. (a) The pipeline of our framework to tackle the task of human de-occlusion, which contains two stages of mask completion and content recovery. (b) Some applications which demonstrate better results can be obtained after human de-occlusion.

In this paper, we aim at the problem of estimating the invisible masks and the appearance content for humans. We refer to such a task as *human de-occlusion*. Compared with general amodal perception, human de-occlusion is a more special and important task, since completing human body plays key roles in many vision tasks, such as portrait editing [1], 3D mesh generation [19], and pose estimation [11] as illustrated in Fig 1 (b). Different from common object de-occlusion task (*e.g.*, vehicle, building, furniture), human de-occlusion presents new challenges in three aspects. First, humans are non-rigid bodies and their poses vary dramatically. Second, the context relations between humans and backgrounds are relatively inferior to common objects which are usually limited to specific scenes. Third, to segment and recover the invisible portions of humans, the algorithm should be able to recognize the occluded body parts to be aware of symmetrical or interrelated patches.

To precisely generate the invisible masks and the appearance content of humans, we propose a two-stage framework to accomplish human de-occlusion as shown in Fig 1 (a). The first stage segments the invisible portions of humans,

*The work was mainly done during an internship at ByteDance Inc.
†Corresponding author.

and the second stage recovers the content inside the regions obtained in the previous stage. In the testing phase, the two stages are cascaded into an end-to-end manner.

Previous methods [54, 6, 50] adopt perfect modal [1] mask as an extra input (*e.g.*, ground-truth annotation or well-segmented mask), and expect to output amodal mask. However, obtaining ground-truth or accurate segmented masks is non-trivial in actual applications. To address the issue, our first stage utilizes a stacked hourglass network [31] to segment the invisible regions progressively. Specifically, the network first applies a hourglass module to refine the inaccurate input modal mask, then a secondary hourglass module is applied to estimate the integrated amodal mask. In addition, our network benefits from human parsing and typical pose masks. Inspired by [22, 50], human parsing pseudo labels are served as auxiliary supervisions to bring prior cues and typical poses are introduced as references.

In the second stage, our model recovers the appearance content inside the invisible portions. Different from typical inpainting methods [36, 27, 52, 30, 49], the context information inside humans should be sufficiently explored. To this end, a novel parsing guided attention (PGA) module is proposed to capture and consolidate the context information from the visible portions to recover the missing content across multiple scales. Briefly, the module contains two attention streams. The first path isolates different body parts to obtain relations of the missing parts with other parts. The second spatially calculates the relationship between the invisible and the visible portions to capture contextual information. Then the two streams are fused by concatenation and the module outputs an enhanced deep feature. The module works at different scales for stronger recovery capability and better performance.

As most existing amodal perception datasets [59, 37, 16] focus on amodal segmentation and few human images are involved, an amodal perception dataset specified for human category is required to evaluate our method. On account of this, we introduce an Amodal Human Perception dataset, namely AHP. There are three main advantages of our dataset: **a)** the number of humans in AHP is comparatively larger than other amodal perception datasets; **b)** the occlusion cases synthesized from AHP own amodal segmentation and appearance content ground-truths from real-world scenes, hence subjective judgements and consistency of the invisible portions from individual annotators are unnecessary; **c)** the occlusion cases with expected occlusion distribution can be readily obtained. Existing amodal perception datasets have fixed occlusion distributions but some scenarios may have gaps with them.

Our contributions are summarized as follows:

- We propose a two-stage framework for precisely generating the invisible parts of humans. To the best of our knowledge, this is the first study which both considers human mask completion and content recovery.
- A stacked network structure is devised to do mask completion progressively, which specifically refines inaccurate modal masks. Additionally, human parsing and typical poses are introduced to bring prior cues.
- We propose a novel parsing guided attention (PGA) module to capture body parts guidance and context information across multiple scales.
- A dataset, namely AHP, is collected to settle the task of human de-occlusion.

[1] The terms 'modal' and 'amodal' are used to refer to the visible and the integrated portions of an object respectively.

2. Related Work

Amodal Segmentation and De-occlusion. The task of modal segmentation is to assign categorical or instance label for each pixel in an image, including semantic segmentation [28, 4, 57, 33] and instance segmentation [38, 13, 3]. Differently, amodal segmentation aims at segmenting the visible and estimating the invisible regions, which is equivalent to the amodal mask, for each instance. Research on amodal segmentation emerges from [25] which iteratively enlarges detected box and recomputes heatmap of each instance based on modally annotated data. Afterwards, AmodalMask [59], MLC [37], ORCNN [8] and PCNets [54] push forward the field of amodal segmentation by releasing datasets or techniques.

De-occlusion [54, 6, 50] targets at predicting the invisible content of instances. It is different from general inpainting [27, 52, 30, 49] which recovers the missing areas (manual holes) to make the recovered images look reasonable. At present, de-occlusion usually depends on the invisible masks predicted by amodal segmentation. SeGAN [6] studies amodal segmentation and de-occlusion of indoor objects with synthetic data. Yan *et al*. [50] propose an iterative framework to complete cars. Xu *et al*. [48] aim at portrait completion without involving amodal segmentation. Compared with them, we jointly tackle the tasks of human amodal segmentation and de-occlusion. Not only that, humans commonly have larger variations in shapes and texture details than rigid vehicles and furniture.

Amodal Perception Datasets. Amodal perception is a challenging task [8, 37, 59, 6] aiming at recognizing occluded parts of instances. Zhu *et al*. [59] point out that humans are able to predict the occluded regions with high degrees of consistency though the task is ill-posed. There are some pioneers building amodal segmentation datasets. COCOA [59] elaborately annotates modal and amodal masks of $5,000$ images in total which originates from COCO [26]. To alleviate the problem that the data scale of COCOA is too small, Qi *et al*. [37] establish KINS with $14,991$ images

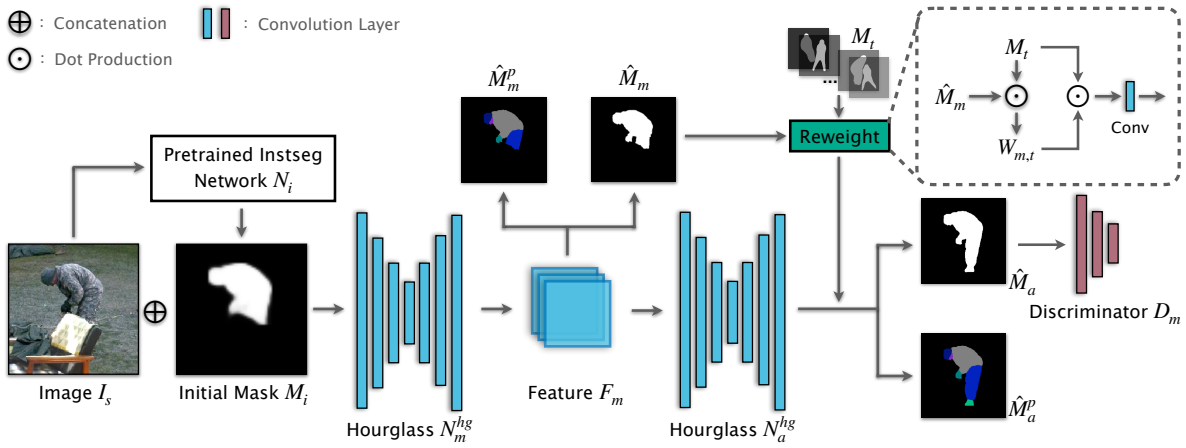

Figure 2. **Illustration of the refined mask completion network.** A pretrained instance segmentation network N_i is applied to obtain the initial modal mask M_i from the input image I_s which contains an occluded human. Then one hourglass module N_m^{hg} outputs a refined modal mask $\hat{M}_m$ and the corresponding parsing result $\hat{M}_m^p$. To complete the modal mask, another hourglass module N_a^{hg} is stacked with the template masks which served as prior appearance cues and it finally outputs the amodal mask $\hat{M}_a$ and the parsing result $\hat{M}_p^a$. At the end, a discriminator D_m is applied to improve the quality of the generated amodal mask.

for two main categories of 'people' and 'vehicle'. In [6], Ehsani *et al*. introduce a synthetic indoor dataset namely DYCE. SAIL-VOS [16] is also synthetic but a video dataset by leveraging a game simulator. Our dataset AHP focuses on amodal perception of humans on account of its wide applications. Particularly, the differences with the above datasets mainly lie in two aspects: **a)** our dataset greatly extends the quantity ($\sim$ 56k) of humans in real-world scenes and provides amodal mask annotations; **b)** our dataset emphasizes on completing both invisible masks and appearance content with trustworthy supervisions provided, to overcome the previous learning techniques with only modal supervisions in [54, 8].

Human Related Works. There are extensive literature on humans, such as detection [47, 58, 55], parsing [53, 56, 45, 40] and pose estimation [32, 2, 29, 44] *etc*. Our work attempts to benefit from these studies for better performance.

3. Method

3.1. Overview

We propose a two-stage framework to accomplish human de-occlusion. The first stage segments the invisible portions of humans and the second recovers the appearance content inside, as shown in Fig 2 and Fig 3 respectively.

3.2. Refined Mask Completion Network

The refined mask completion network targets at segmenting the invisible masks. As mentioned in Sec. 1, most methods first take perfect but non-trivial modal masks as input and the predicted amodal masks will be subtracted with them to get the expected invisible masks. Some of them also apply modal masks from instance segmentation methods, but the gap between amodal perception and instance segmentation annotations are not being noticed. To this end, our mask completion network firstly refines trustless initial modal masks. Specifically, an existing instance segmentation network N_i is applied to obtain the initial modal masks M_i from the input image I_s which contains an occluded human. Then the image and the initial mask are concatenated and fed into the first hourglass module N_m^{hg} to obtain a refined binary modal mask $\hat{M}_m$. The supervision on corresponding parsing result $\hat{M}_m^p$ is accompanied to strengthen the semantic understanding of body parts. The modal recognition process can be formulated as:

$$\hat{M}_m,\ \hat{M}_m^p = N_m^{hg}\left(I_s,\ N_i\left(I_s\right)\right). \tag{1}$$

Next, our network needs to complete the refined modal mask. One direct solution is to augment another amodal branch to accomplish it. However, we empirically find the direct solution will significantly degrade the performance of the refined modal mask. We suspect that completing the amodal mask requires analogous feature both in the occluded and the visible portions of the human, which is opposite to the previous modal mask recognition paying attention on the visible regions only. Accordingly, another hourglass module N_a^{hg} is stacked behind the first one to estimate the amodal mask $\hat{M}_a$ and the parsing result $\hat{M}_a^p$.

Since it can be asserted that the target amodal masks are integrated humans, some typical poses can be implanted into the network as prior appearance cues. We collect a batch of template masks by running k-means on the annotations of the training set, denoted as M_t. Then the ℓ_2 distances $D_{m,t}$ of $\hat{M}_m$ with each template mask is calculated and the attention weight vector can be obtained by $W_{m,t} = 1/D_{m,t}$. The vector $W_{m,t}$ is multiplied back

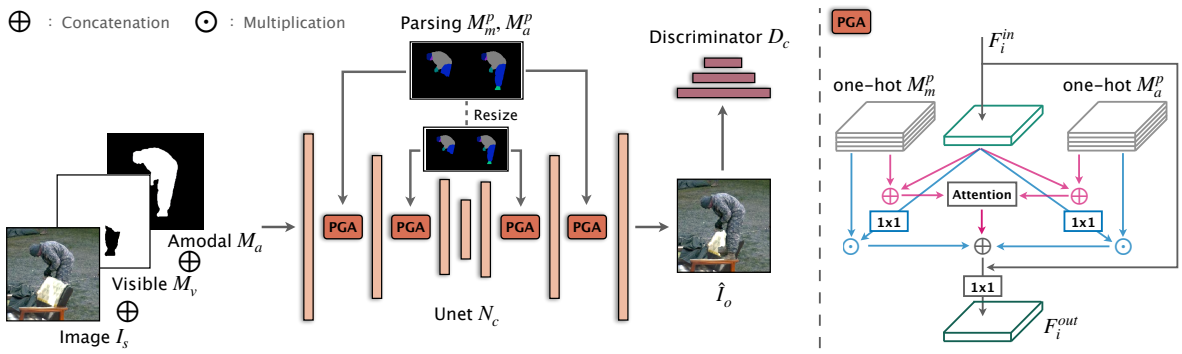

Figure 3. **Illustration of the parsing guided content recovery network. Left**: We adopt Unet [39] with partial convolution [27] as the basic architecture. The image I_s concatenated with the visible mask M_v and the amodal mask M_a is passed into the network N_c to recover content of the invisible portions. To leverage the guidance of body part cues M_m^p and M_a^p from the previous stage, our proposed Parsing Guided Attention (PGA) module enhances deep features at multiple scales. Finally a discriminator D_c is applied to identify the quality of the output image I_o. **Right**: The details of our proposed PGA module which contains two attention streams to capture context relations. The first stream (cyan) decomposes the feature into different body parts and the second (magenta) tries to establish the pixel-level relationship between the visible context and the invisible regions.

with the template masks to highlight suitable candidates and a convolution layer without activation is applied to combine these re-weighted templates as an additional feature before making predictions. The amodal completion process is given by:

$$\hat{M}_a,\ \hat{M}_a^p = N_a^{hg}\left(F_m \oplus \hat{M}_m,\ \text{Conv.}(M_t \odot W_{m,t})\right), \quad (2)$$

where F_m is the intermediate feature from the last hourglass module.

The cross-entropy loss $\mathcal{L}_{CE}(*)$ is applied to supervise the modal and the amodal predictions and the human parsing. The idea in adversarial learning can be borrowed to boost the performance by adopting a discriminator D_m and the perceptual loss [9] $\mathcal{L}_{prec}(*)$ which measures the distances of the extracted features of the generated masks and the ground-truth masks. The several loss functions are formulated as follows:

$$\begin{aligned}
\mathcal{L}_{seg} &= \mathcal{L}_{CE}(\hat{M}_m, M_m) + \mathcal{L}_{CE}(\hat{M}_a, M_a) + \\
&\quad \mathcal{L}_{CE}(\hat{M}_m^p, M_m^p) + \mathcal{L}_{CE}(\hat{M}_a^p, M_a^p), \\
\mathcal{L}_{adv} &= \mathbb{E}_{\hat{M}_a}[\log(1 - D_m(\hat{M}_a))] + \mathbb{E}_{M_a}[\log D_m(M_a)], \\
\mathcal{L}_{gen} &= \mathcal{L}_{\ell 1}(\hat{M}_a, M_a) + \mathcal{L}_{prec}(\hat{M}_a, M_a),
\end{aligned} \quad (3)$$

where $\mathcal{L}_{\ell 1}(*)$ denotes the reconstruction loss. Then the final loss can be summarized with proper coefficients:

$$\mathcal{L}_m = \lambda_1 \mathcal{L}_{seg} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{gen}. \quad (4)$$

3.3. Parsing Guided Content Recovery Network

The mask completion network is able to localize the invisible regions and the body parts by subtracting the refined modal results from the amodal ones, and the goal of this stage is to recover the appearance content inside. Different from general inpainting methods, the missing content is some parts of a human instead of other surroundings to make the recovered images look plausible. In spite of this distinction, some well studied methods can be referred and we adopt Unet [39] with partial convolution [27] as our architecture, as shown on the left of Fig. 3.

Our pipeline is straightforward. At first, the visible mask M_v and the amodal mask M_a are concatenated with the image I_s containing an occluded human as input. The visible mask tells the network N_c which pixels need to be recovered and the amodal mask points out where the integrated human occupies in the image. Additionally, the body part cues M_m^p and M_a^p from the previous stage are aggregated into the deep features by our proposed Parsing Guided Attention (PGA) module at multiple scales. Finally the network outputs a human recovered image $\hat{I}_o$ and a discriminator D_c is applied to estimate the quality of the image $\hat{I}_o$. This stage can be formulated as:

$$\hat{I}_o = N_c\left(I_s \oplus M_v \oplus M_a,\ M_m^p,\ M_a^p\right). \quad (5)$$

The PGA module is depicted on the right of Fig 3. Take some scale i for example, the module takes in deep feature F_i^{in} and the parsing masks M_m^p and M_a^p which are converted into two one-hot logits and resized to the same spatial size with F_i^{in}. It contains two attention streams. The first stream (cyan) decomposes the feature into different body parts and compare them. Specifically, the feature F_i^{in} is reduced to the same channel number with the parsing logits (*i.e.* 19), and it is element-wisely multiplied with the two logits to distribute the feature in different body parts. Then the two distributed features are concatenated and a 1×1 convolution layer is applied to yield a feature of useful body parts. The second stream (magenta) tries to estab-

lish the pixel-level relationship between the visible context and the invisible regions. The difference with the self attention [46] is that self attention wildly calculates pixel-wise relations regardless of the semantics of pixels. In particular, we concatenate the two logits behind the input feature, and two 1×1 convolution layers $\phi(*)$ and $\psi(*)$ are followed to extract key features for the visible and the amodal regions respectively, denoted as $K_{vis} = \phi(F_i^{in} \oplus M_m^p)$ and $K_{amo} = \psi(F_i^{in} \oplus M_a^p)$. Then a relationship matrix R can be obtained by:

$$\begin{aligned} \tilde{R} &= (M_v \odot K_{vis})^T \left((1 - M_v) \odot K_{amo}\right); \\ R &= \text{Softmax}(\tilde{R},\, \text{dim} = 0) \in \mathbb{R}^{HW \times HW}, \end{aligned} \tag{6}$$

where H and W denote the height and width viewed as collapsible spatial dimensions. The matrix R means that given a point of the invisible regions, it gives the pair-wised relevance of the point with the points of all visible regions. Since the relationship matrix contains value for each point, a matrix multiplication of R and the input feature can be applied to extract related information from the visible. At last, the outputs of the two streams are concatenated with the input feature F_i^{in}, and a 1×1 convolution layer is followed to reduce the channel number same with F_i^{in}.

The loss function to optimize our content recovery network is formulated as follows:

$$\begin{aligned} \mathcal{L}_c \; = \beta_1 \, (\mathbb{E}_{\hat{I}_o}[\log(1 - D_c(\hat{I}_o))] + \mathbb{E}_{I_o}[\log D_c(I_o)]) + \\ \beta_2 \, \mathcal{L}_{\ell 1}(\hat{I}_o,\, I_o) + \beta_3 \, \mathcal{L}_{prec}(\hat{I}_o,\, I_o) + \\ \beta_4 \, \mathcal{L}_{style}(\hat{I}_o,\, I_o), \end{aligned} \tag{7}$$

where $\mathcal{L}_{style}(*)$ denotes the style loss proposed in [27].

4. The Amodal Human Perception Dataset

In this section, we describe how we collect human images and obtain their annotations with minimal manual effort. Our method adopts a straightforward pipeline and takes advantage of existing works on human segmentation. As a result, the proposed Amodal Human Perception dataset, namely AHP, contains unoccluded human images with masks. Therefore the occlusion cases can be synthesized by pasting occluders from other datasets onto humans. Moreover, some informative statistics are summarized to analyze our dataset.

4.1. Data Collection and Filtering

Instead of collecting images of people from the Internet, we capitalize on several large instance segmentation and detection datasets to acquire human images. Then a segmentation model is applied to obtain masks for the humans with only box-level annotations. The reason to establish pixel-level annotations is that they can be utilized for occlusion synthesis. Finally, an image filtering scheme with manual effort is applied to construct our dataset. The detailed process is as follows.

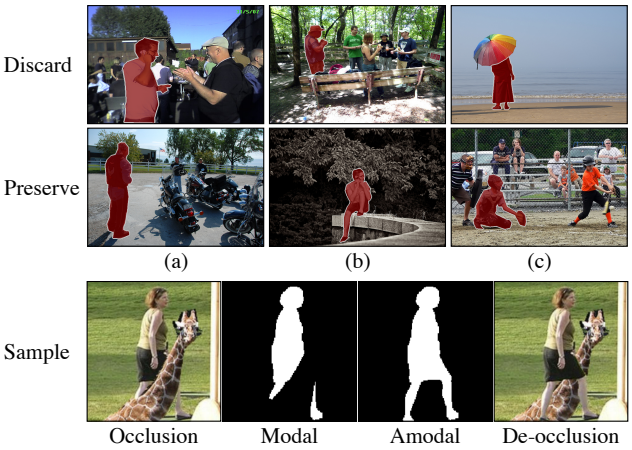

Figure 4. The first two rows show some 'Discard' and 'Preserve' exemplars and the last row demonstrates a synthesized occlusion image with ground-truths of the modal mask and the amodal mask and the final de-occlusion image.

(1) Image Acquisition. We collect human images from several large-scale instance segmentation and detection datasets, including COCO [26], VOC [7] (with SBD [12]), LIP [11], Objects365 [41] and OpenImages [23]. Since each instance in these datasets has been annotated with category label, we simply keep the instances labeled with 'human' for further processing. In addition, human instances that are too small (*e.g.* < 300 pixels) or have overlaps within the 'gutter' (*e.g.* 5 pixels) along four image boundaries are dropped because parts of them are very likely to be out of view.

(2) Human Segmentation. The two largest datasets of Objects365 and OpenImages provide only box-level annotations, which means a great many human instances lack of pixel-level segmentation masks. Hence, a human segmentation model, *i.e.* Deeplab [5] trained on private massive human data, is applied to obtain high-quality segmentation results. In specific, the box for each instance is enlarged by 3 times in height and width to get a cropped image patch which will be fed into the model. Note we only keep segmentation results inside the original annotated box. Al-

	COCO†	OpenImages	Objects365
Human Count	24,806	314,359	669,166
Preserved Count	5,940	17,677	32,982
Preserved Ratio	22.1%	5.62%	4.93%

Table 1. Statistics of the human and the preserved counts. COCO† means COCO unites other datasets, *e.g.* VOC.

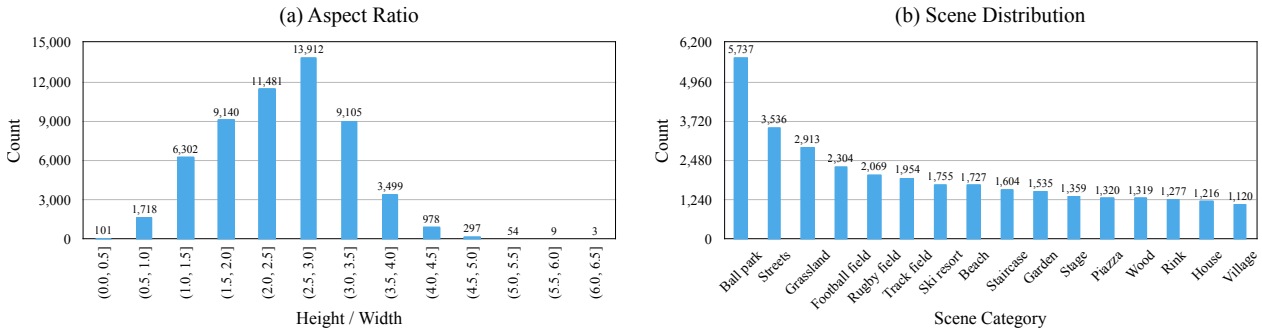

Figure 5. Chart (a) shows the distribution of aspect ratios and chart (b) demonstrates the distribution of scene categories in our dataset.

though the segmentation algorithm may fail on complicated cases, the issue will be addressed next.

(3) Filtering Scheme. Since we aim at collecting unoccluded human instances with accurate masks, there is a need to retain satisfied samples. Three options are set up to route each sample: **Discard**: the human is occluded by other instances (*e.g.* desk, car or other humans) or parts of him/her are out of view; **Preserve**: the human is not occluded and the segmentation is fine; **Refine**: the human is not occluded but the segmentation result is not satisfied. Some 'Discard' and 'Preserve' exemplars are demonstrated in Fig 4. Note for the case of 'Preserve-(c)' that a baseball player wears a glove, we will keep it as the object affects the body structure not too much. But 'Discard-(c)' is an opposite case. There are 7 well-trained annotators filtering the collected human instances. Finally, the number of the 'Preserve' samples reaches over $50,000$ and we think it is already enough to do meaningful research. The samples with the 'Refine' label will be also released for potential exploration.

Once our dataset is built, occlusion training samples can be synthesized. As shown in the last row of Fig 4, an giraffe is cut out from other datasets and pasted onto the woman to form an occlusion case. Since we have the mask annotation of the woman (amodal mask), the ground-truths of the modal mask and the invisible content can be easily inferred.

4.2. Data Statistics

The AHP dataset consists of $56,599$ images following the aforementioned steps. Each sample costs about 3.26 seconds on average. Table 1 shows the total number of the human category and the preserved ratio from each original dataset. It is noticeable that the preserved ratios of OpenImages and Objects365 are significantly smaller than COCO. Besides, the distributions of aspect ratios (height / width) and scene categories are shown in Fig 5. Since the preserved humans are usually standing, it is reasonable that there are nearly half of the samples whose aspect ratios are between 1.5 and 3.5. To estimate the distribution of scene categories, a scene recognition model [43] is applied. Here we only show categories that contain more than $1,000$ samples.

5. Experiments

5.1. Implementation Details

Networks and optimization. The input sizes of the mask completion and the content recovery networks are both 256×256. We adopt CenterMask [24] as our pre-trained instance segmentation model N_i. HRNet [42] is used to generate the pseudo labels of body parts. The hourglass modules of N_m^{hg} and N_a^{hg} and the Unet N_c are inherited from [54] for fair comparison, and they are initialized with random weights. Both D_m and D_c adopt Patch-GAN [17] and share a same structure with four convolution layers. We set $\lambda_1 = \lambda_2 = 1, \lambda_3 = 0.1$ and $\beta_1 = 0.1, \beta_2 = \beta_3 = 1, \beta_4 = 40$ in all experiments. We use SGD with momentum (batch 32, lr $1\mathrm{e}{-3}$, iterations 48k) and Adam [20] (batch 16, lr $1\mathrm{e}{-4}$, iterations 230k) to optimize the completion and the recovery networks respectively. We use PyTorch [35] framework with a NVIDIA Tesla V100 to conduct experiments.

Data splits. One advantage of our AHP dataset is that the distribution of human occlusion ratios can be controlled manually. Based on our observation, we set the probability density of occlusion ratios to $P_{0\sim0.1} = P_{0.1\sim0.2} = P_{0.3\sim0.4} = 1/3$ when training. We draw random instances from COCO [26] to synthesize occlusion cases. At present, our method adopts the simplest synthesis strategy and we leave the problem of how to generate better samples for future research. Our validation set contains 891 images, which is augmented from 297 integrated human samples each with three different occluders. The probability density of occlusion ratios for the validation split is set to $P_{0\sim0.1} = P_{0.1\sim0.2} = P_{0.3\sim0.4} = P_{0.4\sim0.5} = 1/4$ and this set is fixed after generated. To verify the effectiveness of our method in real scenes, several photo editors collect a number of images and they move the foreground instances with similar depths onto the humans manually. Then these artificial occlusion cases are voted by another group of humans and only all passed samples are saved as our test set. The test set contains 56 images due to its steep cost.

Evaluation metrics. To evaluate the quality of the com-

Method	Syn.		Real	
	$\ell_1 \downarrow$	IoU $\uparrow$	$\ell_1 \downarrow$	IoU $\uparrow$
Mask-RCNN [13]	0.2402	78.4/26.9	0.2511	75.6/23.8
Deeplab [5]	0.2087	70.7/20.9	0.2179	75.7/23.5
Pix2Pix [18]	0.2329	69.6/19.2	0.2376	68.0/16.0
SeGAN [6]	0.2545	76.7/23.6	0.2544	77.7/19.0
OVSR [50]	0.1830	80.2/28.1	0.1809	82.9/25.6
PCNets [54]	0.1959	83.1/29.1	0.2218	81.3/31.2
Ours	**0.1500**	**84.6/43.7**	**0.1635**	**86.1/40.3**

Table 2. The comparison results of mask completion task on our AHP dataset. 'Syn.' and 'Real' denote synthesized and real validation images. Our method improves over other techniques both in ℓ_1 error of the amodal masks and IoUs of the amodal and the invisible masks.

pleted masks, we adopt ℓ_1 distance and Intersection over Union (IoU) as our metrics. For the recovered images, we adopt ℓ_1 distance and Fréchet Inception Distance (FID) [15] score which measures the similarity between the ground-truth images and our generated results.

5.2. Results

Quantitative comparison. We compare our method to recent state-of-the-art techniques on our AHP dataset. For the mask completion, some general methods widely applied in other fields like Mask-RCNN [13], Deeplab [5] and Pix2Pix [18] are adopted. There are another group of methods specializing in handling the amodal perception task like SeGAN [6], OVSR [50] and PCNets [54]. Among them, OVSR and PCNets declare they have achieved or surpassed state-of-the-art performance. As shown in Table 2, our method has lower ℓ_1 error and better IoU results on amodal masks both on synthesized and real validation images. It is worth mentioning that we have a significant superiority over the invisible masks. For the task of content recovery, we also selected the two types of general and specialized methods to compare. As shown in Table 3, our method again performs over others on both validation sets.

Qualitative comparison. Fig 6 shows a sample of predicted amodal masks and recovered images on our AHP dataset. For the mask completion task, Pix2Pix [18] and SeGAN [6] have difficulties at completing the occluded humans and their results are not satisfying. Our method has more reasonable and better predicted masks compared to OVSR [50] and PCNets [54]. For the task of content recovery, Deepfillv2 [52] seems blurring and there are apparent artifacts in the results of Pix2Pix [18] and SeGAN [6]. Again our method has better recovery performance compared to OVSR [50] and PCNets [54]. More results are provided in the supplementary material.

Method	Syn.		Real	
	$\ell_1 \downarrow$	FID $\downarrow$	$\ell_1 \downarrow$	FID $\downarrow$
Pix2Pix [18]	0.1126	19.66	0.1031	29.63
Deepfillv2 [52]	0.1127	21.61	0.1026	32.48
SeGAN [6]	0.1122	23.01	0.1027	35.21
OVSR [50]	0.0940	27.15	0.0917	36.23
PCNets [54]	0.0936	18.50	0.0911	28.30
Ours	**0.0519**	**13.85**	**0.0617**	**19.49**

Table 3. The comparison results of content recovery task on our AHP dataset. 'Syn.' and 'Real' denote synthesized and real validation images.

5.3. Ablation Study

To understand our framework further, we conduct extensive experiments on the synthesized and fixed validation images (Sec 5.1) to prove the effectiveness of our method.

Refined mask completion network. Table 4 shows the results. The baseline takes the input image I_s and the initial mask M_i, then outputs the amodal mask $\hat{M}_a$ by networks N_m^{hg} and N_a^{hg} (line 1). The discriminator D_m improves the quality of the amodal mask M_a by 0.3% (line 2). To obtain precise invisible mask, the introduction of the modal segmentation M_m significantly refines the modal mask result by 4.2% but degrades the performance of the amodal by 1.2%. Fortunately the final invisible mask has 4.3% improvements (line 3). It shows that the modal segmentation has negative effects on the amodal completion, and we speculate obtaining the amodal mask requires analogous feature both in the occluded and visible portions of the humans. To solve the problem, the accompanying parsing branches (M_m^p, M_a^p) are leveraged to bring in extra semantic guidance and template pose masks are utilized. The parsing improves 2.3% and 0.8% for the modal and amodal tasks respectively (line 4) and the network benefits from the templates by 0.7% and 0.2% (line 5). It is noticeable that the IoU metric of the invisible mask boosts 5.4% and 3.9% of the two proposed modules. Finally, after we unite these

	Discriminator	Modal	Parsing	Templates	IoU $\uparrow$
1					77.5/84.0/30.3
2	✓				77.5/84.3/30.0
3	✓	✓			81.7/83.1/34.6
4	✓	✓	✓		84.0/83.9/40.0
5	✓	✓		✓	82.4/83.3/38.5
6	✓	✓	✓	✓	**84.7/84.6/43.7**

Table 4. Ablation study of the refined mask completion network. The three columns of each IoU result represent the modal, the amodal and the invisible masks respectively.

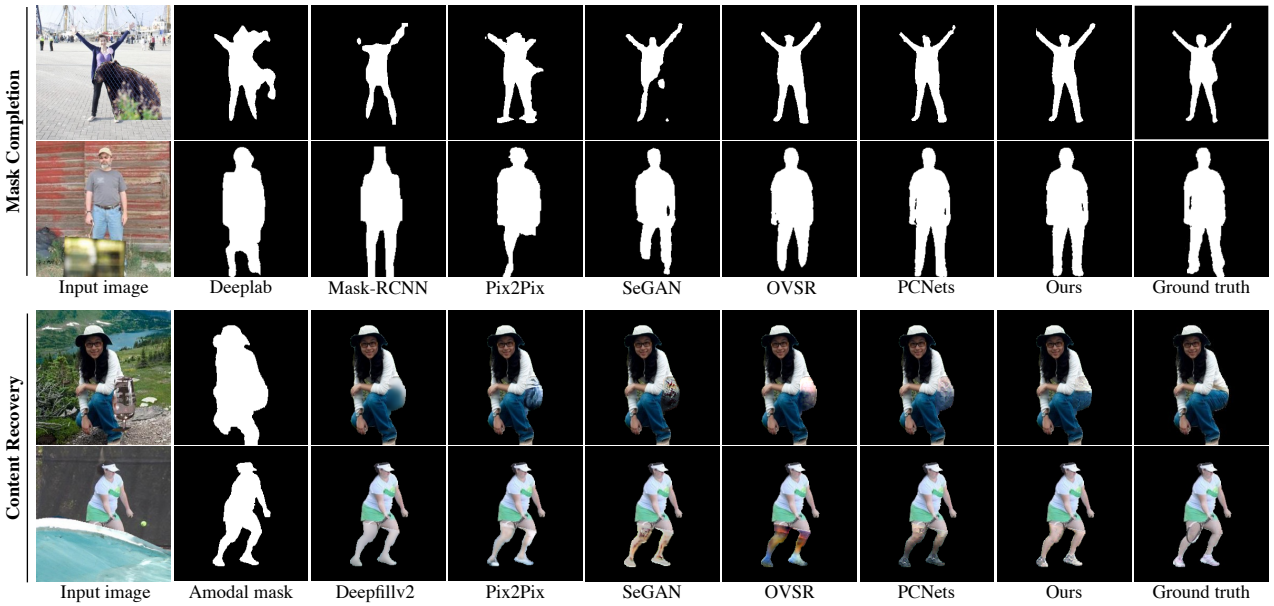

Figure 6. Qualitative comparison results of the mask completion and the content recovery tasks on our AHP dataset.

w	0.0	0.1	0.3	0.5	0.7	0.9	1.0
FID ↓	18.04	18.19	**17.85**	18.74	18.61	19.11	19.71

Table 5. Ablation study of the proportion of the background.

parts together, the IoUs of the modal and invisible masks are remarkably improved by 7.2% and 13.4% respectively compared to the baseline (line 6 vs. line 1).

Parsing guided content recovery network. In Table 5, we analyze the effect of the background first. Intuitively the network should interpolate the invisible portions referring to the visible parts of the human only. However, the background may support context information to aid where the content can be imitated. Therefore, the new input image can be written as: $I_s^{'} = I_s * M_a + I_s * (1 - M_a) * w$, where w is the proportion of the background. The baseline is Unet with partial convolution as [54] with $w = 1$. The experiments show that it gains 1.9 points when $w = 0.3$. Further, our proposed PGA module is disassembled to two attention streams and analyzed as shown in Table 6. The two streams are evaluated individually and the comparison with simply cascading the two streams shows our structure has better performance. Specifically, the first attention stream (Attention.B) separating different body parts boosts the performance by 2.94 points and the second stream (Attention.T) of establishing the relationship between the visible context and the invisible regions gains 2.3 points. Assembling the two streams in a cascade manner yields 5.04 points improvement. Lastly, our proposed structure depicted on the right of Fig 3 further boosts extra 0.8 points.

	Bg (0.3)	Attention.B	Attention.T	Structure	FID ↓
1					19.66
2	✓				17.85
3		✓			16.76
4			✓		17.37
5	✓	✓	✓	Cascade	14.66
6	✓	✓	✓	Fusion	**13.85**

Table 6. Ablation study of the parsing guided content recovery network. 'Bg' denotes the background, and the columns 'Attention.B' and 'Attention.T' correspond to the two attention streams.There are two structures to assemble them: 'Cascade' and 'Fusion'.

6. Conclusion

In this paper, we tackle *human de-occlusion* which is a more special and important task compared with de-occluding general objects. By refining the initial masks from the segmentation model and completing the modal masks, our network is able to precisely predict the invisible regions. Then the content recovery network equipped with our proposed PGA module recovers the appearance details. Our AHP dataset has prominent advantages compared to the current amodal perception datasets. Extended studies on how to generate more realistic samples and aggregate deep features of the two tasks will be explored.

7. Acknowledgements

This work was in part supported by National Natural Science Foundation of China (No. 61876212) and Zhejiang Lab (No. 2019NB0AB02).

References

[1] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 1

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 3

[3] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. BlendMask: Top-down meets bottom-up for instance segmentation. In *IEEE international conference on computer vision*, 2020. 2

[4] 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. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *European conference on computer vision (ECCV)*, pages 801–818, 2018. 5, 7

[6] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 3, 7

[7] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International journal of computer vision*, 88(2):303–338, 2010. 5

[8] Patrick Follmann, Rebecca Kö Nig, Philipp Hä Rtinger, Michael Klostermann, and Tobias Bö Ttger. Learning to see the invisible: End-to-end trainable amodal instance segmentation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1328–1336. IEEE, 2019. 2, 3

[9] Leon A. Gatys, Alexander S. Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 4

[10] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 580–587, 2014. 1

[11] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 1, 5

[12] Bharath Hariharan, Pablo Arbelaez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In *International Conference on Computer Vision*, 2011. 5

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE international conference on computer vision*, pages 2961–2969, 2017. 2, 7

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 1

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, pages 6626–6637, 2017. 7

[16] Y.-T. Hu, H.-S. Chen, K. Hui, J.-B. Huang, and A. G. Schwing. Sail-vos: Semantic amodal instance level video object segmentation – a synthetic dataset and baselines. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 2, 3

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 6

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

[19] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Neural Information Processing Systems*, pages 1097–1105, 2012. 1

[22] Weicheng Kuo, Anelia Angelova, Jitendra Malik, and Tsung-Yi Lin. Shapemask: Learning to segment novel objects by refining shape priors. *IEEE International Conference on Computer Vision*, 2019. 2

[23] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, et al. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv preprint arXiv:1811.00982*, 2018. 5

[24] Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. 2020. 6

[25] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision*, pages 677–693. Springer, 2016. 1, 2

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 5, 6

[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. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 2, 4, 5

[28] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015. 1, 2

[29] Yue Luo, Jimmy Ren, Zhouxia Wang, Wenxiu Sun, Jinshan Pan, Jianbo Liu, Jiahao Pang, and Liang Lin. Lstm pose machines. In *IEEE conference on computer vision and pattern recognition*, pages 5207–5215, 2018. 3

[30] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2

[31] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

[32] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 3

[33] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *IEEE international conference on computer vision*, pages 1520–1528, 2015. 2

[34] Stephen E Palmer. *Vision science: Photons to phenomenology*. MIT press, 1999. 1

[35] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[36] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[37] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3014–3023, 2019. 1, 2

[38] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 1, 2

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[40] Tao Ruan, Ting Liu, Zilong Huang, Yunchao Wei, Shikui Wei, and Yao Zhao. Devil in the details: Towards accurate single and multiple human parsing. In *AAAI Conference on Artificial Intelligence*, volume 33, pages 4814–4821, 2019. 3

[41] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A large-scale, high-quality dataset for object detection. In *IEEE international conference on computer vision*, pages 8430–8439, 2019. 5

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *CVPR*, 2019. 6

[43] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. *arXiv preprint arXiv:1602.07261*, 2016. 6

[44] Manchen Wang, Joseph Tighe, and Davide Modolo. Combining detection and tracking for human pose estimation in videos. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[45] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. Learning compositional neural information fusion for human parsing. In *IEEE International Conference on Computer Vision*, pages 5703–5713, 2019. 3

[46] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 5

[47] Xinlong Wang, Tete Xiao, Yuning Jiang, Shuai Shao, Jian Sun, and Chunhua Shen. Repulsion loss: Detecting pedestrians in a crowd. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7774–7783, 2018. 3

[48] Xian Wu, Rui-Long Li, Fang-Lue Zhang, Jian-Cheng Liu, Jue Wang, Ariel Shamir, and Shi-Min Hu. Deep portrait image completion and extrapolation. *IEEE Transactions on Image Processing*, 29:2344–2355, 2019. 2

[49] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding. Image inpainting with learnable bidirectional attention maps. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019. 2

[50] Xiaosheng Yan, Feigege Wang, Wenxi Liu, Yuanlong Yu, Shengfeng He, and Jia Pan. Visualizing the invisible: Occluded vehicle segmentation and recovery. In *IEEE International Conference on Computer Vision*, pages 7618–7627, 2019. 1, 2, 7

[51] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 1

[52] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. 2, 7

[53] Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. *arXiv preprint arXiv:1909.11065*, 2019. 3

[54] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene deocclusion. In *IEEE conference on computer vision and pattern recognition (CVPR)*, June 2020. 1, 2, 3, 6, 7, 8

[55] Shifeng Zhang, Longyin Wen, Xiao Bian, Zhen Lei, and Stan Z Li. Occlusion-aware r-cnn: detecting pedestrians in a crowd. In *European Conference on Computer Vision*, pages 637–653, 2018. 3

[56] Ziwei Zhang, Chi Su, Liang Zheng, and Xiaodong Xie. Correlating edge, pose with parsing. In *IEEE Conference*

on Computer Vision and Pattern Recognition, pages 8900–8909, 2020. 3

[57] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *IEEE conference on computer vision and pattern recognition*, pages 2881–2890, 2017. 2

[58] Chunluan Zhou and Junsong Yuan. Bi-box regression for pedestrian detection and occlusion estimation. In *European Conference on Computer Vision*, pages 135–151, 2018. 3

[59] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1464–1472, 2017. 1, 2