Article

Removal and Recovery of the Human Invisible Region

Qian Zhang, Qiyao Liang *, Hong Liang and Ying Yang

School of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China; zhangqian8266@163.com (Q.Z.); liangh@upc.edu.cn (H.L.); z19070045@s.upc.edu.cn (Y.Y.)

* Correspondence: z20070082@s.upc.edu.cn

Abstract: The occlusion problem is one of the fundamental problems of computer vision, especially in the case of non-rigid objects with variable shapes and complex backgrounds, such as humans. With the rise of computer vision in recent years, the problem of occlusion has also become increasingly visible in branches such as human pose estimation, where the object of study is a human being. In this paper, we propose a two-stage framework that solves the human de-occlusion problem. The first stage is the amodal completion stage, where a new network structure is designed based on the hourglass network, and a large amount of prior information is obtained from the training set to constrain the model to predict in the correct direction. The second phase is the content recovery phase, where visible guided attention (VGA) is added to the U-Net with a symmetric U-shaped network structure to derive relationships between visible and invisible regions and to capture information between contexts across scales. As a whole, the first stage is the encoding stage, and the second stage is the decoding stage, and the network structure of each stage also consists of encoding and decoding, which is symmetrical overall and locally. To evaluate the proposed approach, we provided a dataset, the human occlusion dataset, which has occluded objects from drilling scenes and synthetic images that are close to reality. Experiments show that the method has high performance in terms of quality and diversity compared to existing methods. It is able to remove occlusions in complex scenes and can be extended to human pose estimation.

Citation: Zhang, Q.; Liang, Q.; Liang, H.; Yang, Y. Removal and Recovery of the Human Invisible Region. *Symmetry* **2022**, *14*, 531. <https://doi.org/10.3390/sym14030531>

Academic Editor: Mihai Postolache

Received: 17 February 2022

Accepted: 2 March 2022

Published: 4 March 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: amodal completion; content recovery; human de-occlusion; human occlusion dataset

1. Introduction

With the development of computer vision, more and more branches have been derived in recent years, including object detection [1–5] and human pose estimation [6–10]. Although all these branches have achieved good results, they still pose a challenge in the case of occlusion. For example, in practical applications in agriculture, robots are generally used to pick and transport crops [11–13]. The principles are all based on the use of target detection techniques in computer vision, and the crops are very easily obscured during the picking process, making production much less efficient. In [14], although depth cameras were incorporated for depth measurement, no satisfactory results were achieved in terms of detection speed. In industrial production, target detection techniques have been incorporated into many operating scenarios in recent years to prevent major accidents. However, operational scenarios are often complex, and workers are highly susceptible to being obscured by surrounding buildings, making human nodes not accurately predictable and not the first warning in the event of a breach, resulting in some safety concerns remaining. Researchers have also optimized processes by incorporating various mechanisms, such as in target detection [15], where instead of predicting a single instance for each candidate frame, a set of potentially highly overlapping instances is predicted. For human pose estimation, ref. [16] proposed instance cues and recurrent refinement. For the case of two targets in a target box, each is fed into the network twice using the instance cue corresponding to the respective target. Although both achieved good results, they still could not be completely solved.

Therefore, for this type of problem, another branch has emerged—image de-occlusion. Image de-occlusion can be seen as a form of image inpainting. Early image inpainting was based on mathematical and physical theories [17] and was accomplished by building geometric models or using texture synthesis to restore small areas of damage. Although small area restoration could be accomplished, it lacked human-like image comprehension and perception. In large regions of broken images, there are problems such as blurred content and missing semantics. With the rapid development of deep learning, people started to use convolutional neural networks for image restoration. Ref. [18] was the first work on image inpainting with GAN (Generative Adversarial Networks), and more and more work on GAN-based image inpainting followed. In recent years, researchers are no longer satisfied with the single task of image restoration and are gradually combining it with de-masking. In other words, we can think of occlusion as a mutilated region of an image and use image inpainting to recover the missing content. For example, [19–21] all restore the occluded content by predicting the occluded region.

This paper focuses on the problem of human de-occlusion. The techniques involved are segmentation, amodal prediction, and image inpainting. As shown in Figure 1, this framework consists of two stages. The first stage segments the instances of the people and then predicts the complete appearance of the human silhouette through the amodal completion network. The second stage recovers the occluded content via the content recovery network. As a whole, the entire framework has a symmetrical character. Unlike previous work [19,21], our study is on people and faces the following main challenges: (1) people are flexible objects with very variable morphology; (2) people appear in scenes with heterogeneous backgrounds and are highly susceptible to interference; and (3) the human body de-occlusion dataset is scarcely available.

Figure 1. The pipeline of frame for the task of human de-occlusion, which contains amodal completion and content recovery.

To address these three challenges, this paper proposes corresponding solutions. In the first stage, to make the generated amodal masks more realistic, the network used a large number of complete human masks as supervision to make the network generate human silhouettes that are more in line with our intuitive perception. In the second stage, the U-net network with symmetrical structure added a VGA (visible guided attention) module, as shown in Figure 4. The purpose of adding the VGA module is to find the relationship between pixels inside and outside the masked region. By calculating the attention map to capture information about the context between them, the quality of the content recovered in a complex context can be addressed. The key remaining challenge is the selection and production of the dataset. It is generally agreed that the selection of occlusion should ensure that the appearance and size are realistic and the occlusion is more natural. In this paper, we select realistic occluded objects in nature, which are more in line with human visual perception.

Contributions of this paper are summarized as follows:

1. We propose a two-stage framework for removing human occlusion to obtain the mask of the human body and recover the occluded area's content. We are a challenging study of humans with highly variable postures.

2. The results of the amodal mask are refined by the fusion of multiscale features on the hourglass network and the addition of a large amount of a priori information.
3. A new visible guided attention (VGA) module was designed to guide low-level features to recover occlusion content by calculating the attention map between the inside and outside of the occlusion region of the high-level feature map.
4. We have used natural occlusions to produce a human occlusion dataset that better matches the visual perception of the human eye. Based on this dataset, it is demonstrated that our model outperforms other current methods. In addition, the problem of unpredictable occluded joints in human pose estimation is solved.

2. Related Work

Amodal Segmentation: Amodal segmentation has a similar task to modal segmentation in that it attaches a label to each pixel in the image. The difference is that amodal segmentation needs to segment out the masked areas of the modal mask. Ref. [22] is the opening work of amodal segmentation, which is done by iteratively enlarging the bounding box and recomputing its heatmap. SeGAN [19] generates amodal masks by inputting the modal mask and the original image into a residual network. Xiao et al. [23] propose a new model that simulates human occlusion target perception based on visible region features and uses shape priors to predict invisible regions.

Image inpainting with generative adversarial network: Image inpainting is the process of inferring and recovering damaged or missing areas based on the known content of the image. Traditional methods of image inpainting based on mathematical and physical theories, which build geometric models or use texture synthesis to restore small areas of damage, can restore small areas but lack human-like image comprehension and perception. In cases where large areas are missing, there are blurred content and missing semantics.

With the development of generative adversarial networks in recent years, researchers have started to experiment with image inpainting using GAN. Ref. [18] is the first paper on image restoration using generative adversarial networks. The principle was to infer the missing image using the surrounding image information, maintain continuity in content using an Encoder-Decoder structure, and maintain continuity in pixels using a discriminator. Since then, researchers have done a great deal of research based on this work. For example, Yang et al. [24] used the most similar intermediate feature layer correlation in deep classification networks to adjust and match patches to produce high-frequency detail information. Iizuka et al. [25] used both a global discriminator and local discriminator to ensure that the generated images conform to the global. Liu et al. [26] propose a partial convolution for irregular missing regions so that convolution is performed only in the active region, and the invisible mask is iterated and shrunk as the network deepens.

Image de-occlusion: Image de-occlusion is a branch of image inpainting that aims to remove occlusions from the target object and recover the content of the occluded region. Ordinary image restoration takes the location information of the missing region directly as input to the network along with the original image [18,24–28]. In contrast, image de-occlusion inputs an image without any missing information into the network to predict the invisible region and then recover the content of the invisible region. Zhan et al. [20] proposed a framework for self-supervised learning, based on the theory that complete complementation is iterated by multiple partial complements to obtain amodal and recover the content of the invisible region using an existing modal mask. Yan et al. [21] proposed two coupled discriminators and a two-path structure with a shared network to perform the segmentation completion and the appearance recovery iteratively. SeGAN [19] also built a two-stage network for image de-masking, but SeGAN [19] only targeted indoor objects. As with [21], both perform de-occlusion for objects with fixed shapes, whereas our network is designed for non-rigid objects with a highly variable pose, such as humans.

3. Method

3.1. Overview

This section introduces the framework for human de-occlusion, consisting of two phases. The first stage predicts the invisible region and generates an amodal mask. The second stage is to recover the content of the invisible region using the amodal mask. This is shown in Figures 2 and 3. The relationship between the occluded regions and the inner and outer regions. Finally, the quality of the generated image I_o is evaluated using a discriminator.

Figure 2. The amodal completion network. A pre-trained instance segmentation model first predicts the visible human areas mask M_v . Then, a large amount of a priori information is obtained from the dataset and concatenated with the multi-scale fused feature maps to input into the various stages of the decoding process. At last, the quality of the predicted M_a is improved by a discriminator.

Figure 3. The content recovery network. The visible area M_v and the invisible area M_i obtained in the first stage are concatenated with the original image as the input to the second stage. The backbone of the content recovery network is U-Net. The visible guided attention is added to obtain.

3.2. Amodal Completion Network

The amodal completion network aims to segment the mask of the invisible area and combine it with the visible mask to generate the amodal mask. This stage uses an hourglass network structure, but with the difference that this network added four branches. The

low-level features generally capture more local detail, while the higher-level features yield more advanced semantic information. Local fine detail and advanced semantic information can be combined by aggregating the underlying features and the up-sampled higher-level features across layers. Inspired by this, this network performed feature fusion of feature maps of different sizes, as shown in Figure 2. It concatenated them with each layer's feature maps in the decoding stage. Finally, the network outputted the final predicted amodal mask.

It is worth noting that to improve the network's effectiveness in predicting amodal, some typical poses are implanted as prior knowledge into the network. Specifically, we used ℓ_2 distance $D_{m,t}$ between the predicted M_v and each ground truth in the training set M_t . After that, the weights of each training set are output using softmax, which is calculated as follows:

$$W_{m,t} = \frac{\exp\{1/D_{m,t}\}}{\sum_1^N \exp\{1/D_{m,t_i}\}} \quad (1)$$

where N denotes the number of M_t . $W_{m,t} = \{W_{m,t}^1, W_{m,t}^2, W_{m,t}^3, \dots, W_{m,t}^N\}$. Each weight $W_{m,t}$ is multiplied with M_t and finally concatenated with the fused feature map.

Finally, this paper judges the generated M_a 's quality by the discriminator Patch-GAN [29]. Cross-entropy loss is used to supervise the M_v and ground truth. Adversarial loss is used to make the generated sample distribution fit the proper sample distribution. Perceptual loss is used to calculate the distance between each layer is generated by feature maps and the proper feature maps. The following loss functions:

$$L_{amo} = L_{CE}(\hat{M}_v, M_v) + L_{CE}(\hat{M}_a, M_a) \quad (2)$$

$$L_{adv} = E_{\hat{M}_a}[\log(1 - D_m(\hat{M}_a))] + E_{M_a}[\log(D_m(M_a))] \quad (3)$$

$$L_{rec} = L_{l_1}(\hat{M}_a, M_a) + L_{prec}(\hat{M}_a, M_a) \quad (4)$$

Finally, we assigned weights to each loss to get the final loss:

$$L_a = \alpha_1 L_{amo} + \alpha_2 L_{adv} + \alpha_3 L_{rec} \quad (5)$$

3.3. Content Recovery Network

The content recovery network aims to recover the content from the invisible areas predicted in the first stage so that the recovered content is consistent semantically and pixel-wise. The network structure is shown in Figure 3. This phase uses a symmetrically structured U-Net network as the architecture for the content recovery network, using both the global discriminator and the local discriminator to judge the recovered content to ensure that the generated images conform to the global semantics while maximizing the clarity and contrast of the local areas.

First, the input has concatenated the M_v and M_i with the original image into five channels. The invisible mask is obtained by taking the intersection of the invisible regions of the M_a and M_i to let the network know which regions' contents need to be recovered. Inspired by [30], low-level features have richer texture details, high-level features have more abstract semantics, and high-level features can guide the complementation of low-level features level by level. Therefore, this network added the visible guided attention (VGA) module to the skip connection. As shown in Figure 4, it integrates the high-level features with the next-level features to guide the low-level features to complete.

The input to the VGA module consists of two parts, as shown in Figure 4a. One part is the feature map F_l obtained from the low-level features through the skip connection, and the other part is the feature map from the deeper layers of the network. Then, these two parts of the feature map are concated, reducing the dimension by 1×1 convolution. To ensure that the structure in the reconstructed features remains consistent with the context, this module added four more sets of dilated convolutions with different rates for aggregation and finally output the feature map.

Figure 4. Schematic diagram of the structure of VGA (visible guided attention). (a) Visible guided attention. The relational feature map is first obtained for the low-level features inside and outside the visible region. The resulting results are up-sampled with the low-level features and then concatenated with the high-level features. Then it is convolved with four different rates of dilated convolution to output the feature map of the layer; (b) the process of calculating a relational feature map.

The computational flow of the relational feature map is shown in Figure 4b. This step is to find the relationship between the pixels inside and outside the occluded region. The feature maps of the visible and invisible regions are first obtained from M_v and M_i , denoted as $R_{vis} = F^d \otimes M_v$ and $R_{inv} = F^d \otimes M_i$, respectively. Then, the dimensionality is reduced to a one-dimensional vector ($\mathbf{R}^{HW \times 1 \times C}$), and a transpose is performed on R_{inv} followed by a multiplication operation ($\mathbf{R}^{HW \times HW \times C}$). Finally, the final relational feature map ($\mathbf{R}^{H \times W \times C}$) is obtained by multiplying with the reduced dimensional F_d ($\mathbf{R}^{HW \times 1 \times C}$). The overall calculation formula is as follows:

$$R = \text{resize}(\text{Softmax}((F^d \otimes M_v)(F^d \otimes M_i)^T, \text{dim} = 0) \otimes F^d) \in \mathbf{R}^{H \times W \times C} \quad (6)$$

For predicted picture y and ground truth $\hat{y}$, the adversarial loss is defined as:

$$L_{adv} = E_{\hat{y}}[\log(1 - D_m(\hat{y}))] + E_y[\log(D_m(y))] \quad (7)$$

The ℓ_1 loss is defined as:

$$L_{l_1} = \frac{1}{hwc} \sum_{i,j,k} |\hat{y}_{i,j,k} - y_{i,j,k}| \quad (8)$$

The style loss is defined as:

$$L_{style} = \sum_j \frac{1}{C_j H_j W_j} \sum_{h=1}^{H_j} \sum_{w=1}^{W_j} \|\varphi_j(\hat{y}) \cdot \varphi_j^T(\hat{y}) - \varphi_j(y) \cdot \varphi_j^T(y)\|^2 \quad (9)$$

The content loss is defined as:

$$L_{con} = \frac{1}{C_j H_j W_j} \|\varphi_j(\hat{y}) - \varphi_j(y)\|^2 \quad (10)$$

C_j , H_j , W_j is the number of channels, height, and width of the j th layer feature map, respectively. $\varphi(\cdot)$ is a feature map of the output of VGG19 [31], the exact layer of which is given in Section 5. To make the image smoother, we also add TV loss (total variation loss):

The overall loss for the content recovery network is defined as:

$$L_c = \beta_1 L_{adv} + \beta_2 L_{l_1} + \beta_3 L_{style} + \beta_4 L_{con} + \beta_5 L_{tv} \quad (11)$$

4. Human Occlusion Dataset

This section presents the human occlusion dataset, including the data's selection, filtering, and production. The dataset was synthesized using authentic images and natural occlusions to match the human visual perception.

4.1. Data Collection and Filtering

We select images of people from several large public datasets for strength segmentation and target detection, including VOC [32] and ATR [33]. In addition, we collect some portrait images from our drilling dataset. We also obtained occluders from the drilling dataset, including objects such as railings, noticeboards, winches, barrels, etc., which are very relevant to the actual situation.

The VOC and ATR datasets are annotated at the pixel level for each category, so we only needed to filter those labeled "Person". The drilling dataset needs to segment the portraits using the pre-trained segmentation model Yolact [34] and eliminate the images that did not work well. The final number of portraits selected from each dataset is shown in Table 1.

Table 1. Statistics on the number of humans obtained from several datasets.

VOC Train2017 [32]	ATR [33]	Drilling
5644	17,706	13,069

4.2. Data Production

We use the photoshop tool to crop the masks in the Drilling dataset. A total of 100 masks were obtained, which are shown in Figure 5. Then, we performed FLIP_LEFT_RIGHT, FLIP_TOP_BOTTOM, ROTATE_90, ROTATE_180, ROTATE_270 operations on the masks and obtained the results shown in Figure 6. In this way, we obtained 600 occlusions.

Figure 5. Several occluders. (a) railing; (b) winch; (c) barrel.

Afterward, the human occlusion dataset is produced using a streamlined operation. The steps are: (1) Generate a mask for each occlusion. (2) The occlusion is randomly covered on the portrait to generate the occluded RGB image. (3) Generate occluded modal mask and invisible mask. (4) Eliminate images with too many or too few occlusion pixels. As such, this dataset consists of five parts: Amodal Masks, Invisible Masks, Modal Masks, Uncover Image, and Covered Image, as shown in Figure 7.

Figure 6. Occlusions generation operations.

Figure 7. Composition of the human occlusion dataset.

5. Experiments

5.1. Implementation Details

The human occlusion dataset consists of 72,838 images. The proportion of the training set, validation set, and test set is 60%, 20%, and 20%, respectively. The segmentation model was pre-trained using Yolact [34]. The hourglass network and U-Net [35] were used as

the backbone of the amodal completion network and the content recovery network. Both networks use Patch-GAN [29] as the discriminator and relu2_2, relu3_4, relu4_2, relu5_2 of VGG19 [31] as the style loss. The relu1_1, relu2_1, relu3_1, relu4_1, and relu5_1 of VGG19 are used as texture loss. We set $\alpha_1 = \alpha_2 = 1$, $\alpha_3 = 0.1$ and $\beta_1 = 0.1$, $\beta_2 = \beta_4 = 1$, $\beta_3 = 1000$, $\beta_4 = 5 \times 10^{-6}$ in all experiments. Pytorch [36] was the framework for the network, Python version 3.8. We used Adam [37] to optimize the amodal completion and the content recovery networks. For the network's generator, the learning rate is set to 1×10^{-3} , betas = (0.9, 0.99). For the discriminator and perceptron of the network, the learning rate is set to 1×10^{-4} , betas = (0.5, 0.999). The amodal completion network and the content recovery network are set to batch 4 and 8. In total, 200 epochs are iterated on Titan X.

The input image size for both networks is 256×256 . The inputs for the amodal completion network are the original and predicted modal masks and the amodal masks obtained by clustering in the training set. The outputs are the predicted amodal masks. For the content recovery network, the input is the original map, modal masks, and invisible masks connected to a 5-channel map, and the output is a de-occluded RGB image. ℓ_1 distance and mIoU (Mean Intersection over Union) are used as evaluation metrics for the amodal completion network, and ℓ_1 and ℓ_2 distances, as well as the FID score [38], are used to evaluate the similarity of ground truth and generated images.

5.2. Comparison with Existing Methods

We conducted experiments on the human occlusion dataset. For the amodal completion task and the content recovery task, control experiments were performed using SeGAN [19], PCNets [20], and OVSR [21]. These three are currently very advanced models, using two stages to remove the occlusion. SeGAN [19] has a relatively simple structure, with only one discriminator to constrain the generated content, and is only applicable to objects with regular shapes. PCNets [20] are trained unsupervised, without using ground truth for supervision, and the final results are often unsatisfactory. OVSR [21] proposed two coupled discriminators and introduced an auxiliary 3D model pool with a relatively complex structure. However, the object of study is a vehicle, which is relatively fixed in shape and color and does not have substantial deformations.

In order to make the experimental design more reasonable, we do two sets of experiments with synthetic images and authentic images on two stages, respectively, as shown in Tables 2 and 3. Table 2 shows the results of the amodal completion task. It can be seen that these models perform better on authentic images than on synthetic images. SeGAN [19] and PCNets [20] perform worse than our model, with lower ℓ_1 error and better mIoU results. Although the ℓ_1 error is higher on synthetic images than PCNets [20], it reaches the lowest on real images, 0.0183 lower than SeGAN [19]. This also demonstrates the excellent generalization ability of our model on amodal completion. Table 3 shows the results of the content recovery task. We can see that the recovery quality of synthetic images is better than that of authentic images. SeGAN [19] and PCNets [20] have limitations in content recovery. Our method has better recovery performance compared to OVSR [21].

Table 2. Performance comparison of several models on the amodal completion task.

Model	Synthetic Images		Authentic Images	
	mIoU $\uparrow$	ℓ_1 $\downarrow$	mIoU $\uparrow$	ℓ_1 $\downarrow$
SeGAN [19]	0.722	0.0836	0.732	0.0821
PCNets [20]	0.783	0.0718	0.773	0.0729
OVSR [21]	0.826	0.0653	0.836	0.0645
ours	0.802	0.0677	0.82	0.0638

Table 3. Performance comparison of several models on the content recovery task.

Model	Synthetic Images			Authentic Images		
	$\ell_1 \downarrow$	$\ell_2 \downarrow$	FID [26] $\downarrow$	$\ell_1 \downarrow$	$\ell_2 \downarrow$	FID [26] $\downarrow$
SeGAN [19]	0.042	0.0403	28.96	0.0418	0.0398	38.36
PCNets [20]	0.0368	0.0346	26.18	0.0366	0.0349	37.57
OVSr [21]	0.0352	0.0338	22.61	0.0348	0.0326	35.4
ours	0.0343	0.033	20.83	0.0344	0.0324	33.28

Figure 8 shows the results of the proposed method compared to these three models. The first two rows show the effect of amodal completion, and it can be seen that the amodal mask predicted by SeGAN [19] and PCNets [20] is not satisfactory. In contrast, our model and OVSr [21] predict more reasonable results. This also shows that adding a large amount of prior information in the training phase can constrain the model to predict in the correct direction. The last two rows show the effect of content recovery. SeGAN [19] is less effective at color filling and texture generation, and PCNets [20] is not as good at texture generation. In contrast, OVSr [21] seems more reasonable in these two aspects, but there is still apparent blurring. On the other hand, our model outperforms all three models in both color filling and texture generation, which demonstrates that the proposed VGA module plays a significant role in content recovery.

Figure 8. Performance comparison of several methods for amodal completion and content recovery tasks.

5.3. Ablation Study

In order to demonstrate the validity of the proposed model, we have done multiple sets of ablation experiments on the various mechanisms of the model.

Amodal Completion Network: Table 4 shows the results of the experiments on the amodal completion network. From the second row, the discriminator improves the results by 3.4%. From the third and fifth rows, it can be seen that adding prior information improves the results by a significant 5.3%. This indicates that a large amount of prior

knowledge constrains the prediction results of the model in a positive direction. From the fourth and fifth rows, perceptual loss improves the results by approximately 2%.

Table 4. Ablation study of the amodal completion network.

	Discriminator	Prior	Perceptual	IoU
1				0.686
2	✓			0.720
3	✓	✓		0.783
4	✓		✓	0.746
5	✓	✓	✓	0.802

Content Recovery Network: In order to keep the non-masked area of the generated image consistent with the original, there is a strategy for the output: $I_o = I \times (1 - M_v) + I_o' \times M_v$. I_o' is the original image that was output. Table 5 shows the results of experiments testing various mechanisms on the VGA. We control experiments on whether the VGA module low-level features and the attention map after up-sampling and high-level features were concatenated or multiplied. In addition, this experiment verified the necessity of dilated convolution. From the results in the first and third rows of the table, it can be seen that the concatenated approach is more effective than the multiplication approach. From the results in the second and fourth rows, it can be seen that the inclusion of the dilated convolution causes the FID to drop by approximately 0.23.

Table 5. Ablation study of the content recovery network.

	Concat	Mul	Dilated Conv	FID
1		✓		24.40
2		✓	✓	24.16
3	✓	✓		22.83
4	✓		✓	22.61

5.4. Human Pose Estimation

The proposed model is a human body de-occlusion model that solves the occlusion problem with the object of study being the human orientation. To demonstrate the effectiveness of the proposed method, we have done several more sets of human pose estimation experiments. Several occlusions are added within a reasonable range of each image of the drilling and VOC dataset with a ratio of 256×256 , respectively. Three human pose estimation models, OpenPose [7], HigherHRNet [6], and AlphaPose [39], were chosen to do comparison experiments with and without occlusion, respectively. As shown in Figure 9, it can be seen that the human body does not predict the invisible joints, or the predicted positions are inaccurate in the occluded case. After removing the occlusion, this model can easily predict the occluded joints. This shows that our model can solve the occlusion problem in the direction of the human subject.

Figure 9. Controlled experiments in occluded and unoccluded situations, respectively. The models used were OpenPose [7], HigherHRNet [6], and AlphaPose [39], respectively.

6. Conclusions

A two-stage framework is proposed to solve the problem of human occlusion as the object of study in computer vision. The first stage predicts the complete contour of the human body and improves the accuracy of the invisible region of the human body by adding a priori information. The second stage incorporates the proposed VGA module to obtain rich multi-scale feature information inside and outside the occluded region and accurately recover the content and texture of the occluded region. Besides, the provided human occlusion dataset is well synthesized and closely resembles the occlusion effect in nature. Experiments show that the proposed model outperforms other models in content generation and texture drawing; however, there is still much scope for optimization in terms of amodal prediction. In addition to this, the proposed method is combined with human pose estimation to solve the problem of unpredictable joint points in occluded regions.

Author Contributions: Conceptualization, Q.Z. and Q.L.; methodology, Q.L.; investigation, Y.Y.; data curation, Q.L.; writing—original draft preparation, Q.L.; writing—review and editing, Q.Z.; visualization, Y.Y.; supervision, Q.Z.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Science Foundation of Shandong Province, grant number ZR2020MF005.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The part of dataset presented in this study are openly available at <https://pan.baidu.com/s/1ESlsJpCtu0EQXVjGC7zHag?pwd=3643> (accessed on 10 February 2022).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
2. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
3. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
4. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [[CrossRef](#)] [[PubMed](#)]
5. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*, 91–99. [[CrossRef](#)] [[PubMed](#)]
6. Cheng, B.; Xiao, B.; Wang, J.; Shi, H.; Huang, T.S.; Zhang, L. Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–20 June 2020; pp. 5386–5395.
7. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
8. Newell, A.; Yang, K.; Deng, J. Stacked hourglass networks for human pose estimation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 483–499.
9. Liu, Z.; Chen, H.; Feng, R.; Wu, S.; Ji, S.; Yang, B.; Wang, X. Deep Dual Consecutive Network for Human Pose Estimation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 525–534.
10. Artacho, B.; Savakis, A. Unipose: Unified human pose estimation in single images and videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–20 June 2020; pp. 7035–7044.
11. Kuznetsova, A.; Maleva, T.; Soloviev, V. Using YOLOv3 algorithm with pre-and post-processing for apple detection in fruit-harvesting robot. *Agronomy* **2020**, *10*, 1016. [[CrossRef](#)]
12. Kamysheva, G.; Osipov, A.; Gataullin, S.; Korchagin, S.; Ignar, S.; Gataullin, T.; Terekhova, N.; Suvorov, S. Artificial Neural Networks and Computer Vision's-Based Phytoindication Systems for Variable Rate Irrigation Improving. *IEEE Access* **2022**, *10*, 8577–8589. [[CrossRef](#)]
13. Korchagin, S.A.; Gataullin, S.T.; Osipov, A.V.; Smirnov, M.V.; Suvorov, S.V.; Serdechnyi, D.V.; Bublikov, K.V. Development of an Optimal Algorithm for Detecting Damaged and Diseased Potato Tubers Moving along a Conveyor Belt Using Computer Vision Systems. *Agronomy* **2021**, *11*, 1980. [[CrossRef](#)]
14. Andriyanov, N.; Khasanshin, I.; Utkin, D.; Gataullin, T.; Ignar, S.; Shumaev, V.; Soloviev, V. Intelligent System for Estimation of the Spatial Position of Apples Based on YOLOv3 and Real Sense Depth Camera D415. *Symmetry* **2022**, *14*, 148. [[CrossRef](#)]
15. Chu, X.; Zheng, A.; Zhang, X.; Sun, J. Detection in crowded scenes: One proposal, multiple predictions. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–20 June 2020; pp. 12214–12223.
16. Dai, H.; Zhou, L.; Zhang, F.; Zhang, Z.; Hu, H.; Zhu, X.; Ye, M. Joint COCO and Mapillary Workshop at ICCV 2019 Keypoint Detection Challenge Track Technical Report: Distribution—Aware Coordinate Representation for Human Pose Estimation. *arXiv* **2020**, arXiv:2003.07232.
17. Barnes, C.; Shechtman, E.; Finkelstein, A.; Goldman, D.B. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.* **2009**, *28*, 24. [[CrossRef](#)]
18. Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; Efros, A.A. Context encoders: Feature learning by inpainting. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2536–2544.
19. Ehsani, K.; Roozbeh, M.; Ali, F. Segan: Segmenting and generating the invisible. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018.
20. Zhan, X.; Pan, X.; Dai, B.; Liu, Z.; Lin, D.; Loy, C.C. Self-supervised scene de-occlusion. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–20 June 2020; pp. 3784–3792.
21. Yan, X.; Wang, F.; Liu, W.; Yu, Y.; He, S.; Pan, J. Visualizing the invisible: Occluded vehicle segmentation and recovery. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 7618–7627.
22. Li, K.; Malik, J. Amodal instance segmentation. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 8–16 October 2016; pp. 677–693.
23. Xiao, Y.; Xu, Y.; Zhong, Z.; Luo, W.; Li, J.; Gao, S. Amodal Segmentation Based on Visible Region Segmentation and Shape Prior. *arXiv* **2020**, arXiv:2012.05598.
24. Yang, C.; Lu, X.; Lin, Z.; Shechtman, E.; Wang, O.; Li, H. High-resolution image inpainting using multi-scale neural patch synthesis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6721–6729.
25. Iizuka, S.; Simo-Serra, E.; Ishikawa, H. Globally and locally consistent image completion. *ACM Trans. Graph.* **2017**, *36*, 1–14. [[CrossRef](#)]
26. Liu, G.; Reda, F.A.; Shih, K.J.; Wang, T.C.; Tao, A.; Catanzaro, B. Image inpainting for irregular holes using partial convolutions. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 85–100.
27. Liu, H.; Wan, Z.; Huang, W.; Song, Y.; Han, X.; Liao, J. PD-GAN: Probabilistic Diverse GAN for Image Inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 9371–9381.

28. Nazeri, K.; Ng, E.; Joseph, T.; Qureshi, F.Z.; Ebrahimi, M. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv* **2019**, arXiv:1901.00212.
29. Isola, P.; Zhu, J.Y.; Zhou, T.; Efros, A.A. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1125–1134.
30. Zeng, Y.; Fu, J.; Chao, H.; Guo, B. Learning pyramid-context encoder network for high-quality image inpainting. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 1486–1494.
31. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
32. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [[CrossRef](#)]
33. Yamaguchi, K.; Kiapour, M.H.; Ortiz, L.E.; Berg, T.L. Parsing clothing in fashion photographs. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3570–3577.
34. Bolya, D.; Zhou, C.; Xiao, F.; Lee, Y.J. Yolact: Real-time instance segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 9157–9166.
35. Ronneberger, O.; Fischer, P.; Brox, T. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 5–9 October 2015; pp. 234–241.
36. Paszke, A.; Gross, S.; Chintala, S.; Chanan, G.; Yang, E.; DeVito, Z.; Lerer, A.; Lin, Z.; Desmaison, A.; Antiga, L. Automatic differentiation in pytorch. In Proceedings of the NIPS 2017 Autodiff Workshop, Long Beach, CA, USA, 9 December 2017.
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; Hochreiter, S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
39. Fang, H.S.; Lu, G.; Fang, X.; Xie, J.; Tai, Y.W.; Lu, C. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv* **2018**, arXiv:1805.04310.