PHAC: Promptable Human Amodal Completion

Seung Young Noh Ju Yong Chang
Kwangwoon University
{kelvinnoh,jychang}@kw.ac.kr

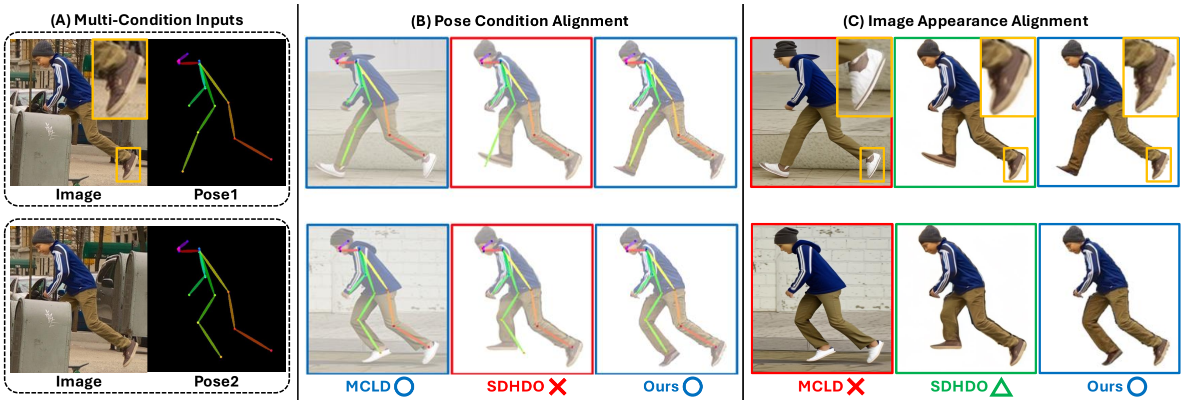

Figure 1. Assume that, as in (A), a single input image is paired with multiple target poses. The desired behavior is that each generated image aligns with its target pose while preserving the visible appearance. However, prior human amodal completion (HAC) methods such as SDHDO [39] often fail to align with the specified pose, as shown in (B), despite leveraging pose information. Pose-guided person image synthesis (PGPIS) methods such as MCLD [28] align well with the pose condition but degrade the visible appearance of the input, especially around the shoes, as shown in (C). While SDHDO preserves the visible appearance better than MCLD, its results remain blurry and show noticeable degradation. In contrast, given the multi-condition inputs in (A), our method simultaneously aligns with the target pose and preserves the visible appearance, producing the user-intended human image.

Abstract

Conditional image generation methods are increasingly used in human-centric applications, yet existing human amodal completion (HAC) models offer users limited control over the completed content. Given an occluded person image, they hallucinate invisible regions while preserving visible ones, but cannot reliably incorporate user-specified constraints such as a desired pose or spatial extent. As a result, users often resort to repeatedly sampling the model until they obtain a satisfactory output. Pose-guided person image synthesis (PGPIS) methods allow explicit pose conditioning, but frequently fail to preserve the instance-specific visible appearance and tend to be biased toward the training distribution, even when built on strong diffusion model priors. To address these limitations, we introduce promptable human amodal completion (PHAC), a new task that completes occluded human images while satisfying both visible appearance constraints and multiple user prompts. Users provide simple point-based prompts, such as additional joints for the target pose or bounding boxes for desired regions; these prompts are encoded using ControlNet modules specialized for each prompt type. These modules inject the prompt signals into a pre-trained diffusion model, and we fine-tune only the cross-attention blocks to obtain strong prompt alignment without degrading the underlying generative prior. To further preserve visible content, we propose an inpainting-based refinement module that starts from a slightly noised coarse completion, faithfully preserves the visible regions, and ensures seamless blending at occlusion boundaries. Extensive experiments on the HAC and PGPIS benchmarks show that our approach yields more physically plausible and higher-quality completions, while significantly improving prompt alignment compared with existing amodal completion and pose-guided synthesis methods.

1. Introduction

Image generation has been widely adopted in recent years and continues to attract strong interest. In particular, conditional image generation, which synthesizes images that satisfy user-specified conditions, has attracted increasing attention relative to unconditional generation, which produces diverse images without explicit guidance. For conditional image generation, beyond output quality and realism, it is crucial that models be easily controlled with diverse user prompts and that the generated results faithfully align with these prompts. In practice, users frequently refine and share text prompts to better express their intended conditions, and frameworks such as ControlNet [62] are widely used to handle diverse conditioning inputs.

Human amodal completion (HAC) takes an image of an occluded human as input and synthesizes a plausible human shape and appearance in the occluded regions while preserving the visible content. Therefore, HAC can be viewed as a form of conditional image generation. Producing a complete human image from an incomplete input enables downstream applications such as novel-view synthesis, 3D avatar generation, and user-guided photo editing. For these reasons, amodal completion, including HAC, remains an active research topic. However, existing approaches [7, 30, 34, 39–41, 53, 55, 59, 67] generally cannot incorporate user-specified prompts as conditioning signals. For example, when a user provides a desired pose for an occluded person, there is no straightforward way to enforce that pose during generation. As a result, users may need to run the model multiple times to obtain a satisfactory result, and models often produce implausible human completions in the process. Recent works [39, 40] can inject pose conditions using pre-trained ControlNet modules [62]. However the generated images often remain poorly aligned with the given pose constraint, as shown in Fig. 1(B).

A promising direction for addressing these limitations is pose-guided person image synthesis (PGPIS). PGPIS takes a source image, a source pose, and a target pose as input, and synthesizes an image that aligns with the target pose while remaining consistent with the source appearance. In this regard, PGPIS can be viewed as conditional image generation, where the source image provides appearance information and the target pose serves as the pose condition. In the HAC setting, the target pose contains all visible joints from the source pose while adding only the invisible ones. The target image should preserve the visible regions while satisfying this extended pose condition. Despite leveraging the strong priors of pre-trained diffusion models (DMs) [14, 47], existing PGPIS methods [3, 12, 28, 31] often struggle to maintain instance-specific visible appearance and instead tend to produce content reflecting the training distribution, as illustrated in Fig. 1(C).

Both amodal completion and PGPIS require synthesizing plausible human appearance in invisible regions based on the visible context. Recent approaches improve performance by leveraging DM priors or fine-tuning the models. However, latent space denoising tends to discard fine-grained visible details. To mitigate this issue, previous work has fine-tuned decoders to reduce information loss in visible regions [1, 39] or has used UV coordinate-based texture maps [11] to preserve appearance [12, 28]. Nevertheless, decoder fine-tuning may introduce blurry outputs and blending artifacts at mask boundaries, while texture map approaches often lose details when UV coordinates are noisy.

To overcome these limitations, we introduce a new task called *promptable human amodal completion (PHAC)*. The goal is to generate human images that satisfy both the visible appearance constraints and multiple user-specified prompts, in which users can specify prompts using only a few points. For pose prompts, the user specifies additional joints for the desired pose given the visible ones. For bounding-box (bbox) prompts, the user simply selects two points. We treat these simple prompts, together with the occluded input image, as conditioning signals for HAC. Consequently, unlike prior HAC [39, 67] and general amodal completion [30, 40, 53, 59] methods, our generated images can reflect user-specified constraints. Beyond steering generation toward user intention, prompts also provide auxiliary guidance for the invisible regions, which enables the diffusion model to produce more physically plausible and higher-quality results. The prompts are injected into the denoising U-Net via ControlNet modules specialized for each prompt type. To preserve the rich priors of pre-trained DMs while achieving strong prompt alignment, we fine-tune only the cross-attention blocks [13].

While the generated human image aligns with the user prompts, the visible appearance can still degrade during latent space denoising and variational autoencoder (VAE) [22] reconstruction. To address this issue, we refine the generated image using an inpainting model [42]. Standard inpainting models synthesize pixels within a masked region and blend them into the surrounding context. By instead using a model that preserves unmasked regions and seamlessly fuses the synthesized content, our refinement step maintains the visible appearance and yields a coherent final image. The refinement module takes the coarse completion, the incomplete image, and the invisible-region mask as inputs, preserves the visible regions exactly, and applies only a small amount of denoising to ensure smooth blending. Denoising from a low noise level, rather than from pure random noise, further helps retain both visible and generated details and preserves boundary consistency.

The key contributions of this work are as follows:

- We introduce the new task of promptable human amodal completion, which completes occluded human images while satisfying both visible appearance constraints and user-specified prompts. We also present the first framework to address this task.
- We propose an inpainting-based refinement module that preserves the visible appearance and synthesized content while maintaining boundary continuity. This module can also serve as a plug-and-play component for improving the outputs of other diffusion models.

- We demonstrate, through extensive quantitative and qualitative evaluations, that our approach achieves superior prompt alignment and produces more physically plausible and higher-quality images than existing amodal completion and PGPIS methods.

2. Related Work

2.1. Conditional Image Generation

Conditional image generation aims to synthesize images that satisfy specified conditions. Early work explored conditioning mechanisms in VAEs [2, 50] and generative adversarial networks (GANs) [10, 36]. Subsequent GAN-based methods extended conditioning to image-to-image translation [18, 68] and text-guided image generation [54, 60]. DMs [14, 47] further advanced conditional synthesis, particularly in large-scale text-to-image generation [38, 45, 48]. However, text prompts alone often fail to express fine spatial constraints such as geometry or human pose. ControlNet [62] addresses this limitation by injecting image-based structural cues (e.g., 2D poses and normal maps) into a pre-trained DM. While ControlNet improves spatial controllability, faithfully preserving the detailed appearance of a specific person remains challenging.

2.2. Amodal Completion

Amodal completion aims to recover both the amodal region and RGB content of occluded areas. Early work predicted only the amodal region, using bounding boxes [15, 20] or amodal masks [5, 9, 16, 21, 25, 26, 43, 46, 51, 65, 69], and subsequent studies inferred amodal masks directly from visible masks [37, 58]. Later two-stage approaches [7, 41, 55, 59] first reconstructed the occluded region and then synthesized its RGB appearance. Although effective within training distributions, these methods often struggle to generalize to complex or unseen occlusion patterns. To improve robustness, prior work restricted the problem to specific categories such as hands [34] or full-body humans [39, 67], and diffusion-based methods [40, 53] leveraged pre-trained generative priors to enhance zero-shot performance. However, existing methods do not support user-specified conditioning and may still lose fine visible details due to denoising.

2.3. Pose-Guided Person Image Synthesis

PGPIS methods synthesize images that align with a target pose while preserving the source appearance. Early PGPIS approaches relied on GANs [6, 32, 33, 49, 61], while later methods improved pose-texture interaction using attention mechanisms [70] and transformer-based architectures [63]. More recent diffusion-based approaches [3, 12, 28, 31] inject the source appearance and target pose into the latent space of pre-trained DMs. Nonetheless, these models are often biased toward training datasets such as DeepFashion [29], and may synthesize appearance that reflects dataset-specific biases rather than preserving the instance-specific visual characteristics of zero-shot inputs.

2.4. Prompts and Human Interaction

User prompting provides an effective mechanism for steering model behavior when automatic inference does not align with user intention. In computer vision, SAM [23] supports point- or box-based prompts to generate segmentation masks, while text-to-image DMs such as Stable Diffusion [47] accept text prompts and structural conditions via ControlNet [62]. Other recent works utilize spatial or semantic prompts for controllable generation, including semantic palettes [24], spatial cues for mesh or image generation [17, 19], and auxiliary prompts that improve tasks such as human mesh recovery [52], pose estimation [56], and 3D object detection [64]. These studies demonstrate the effectiveness of lightweight prompting for guiding complex models. However, to the best of our knowledge, existing approaches neither address amodal completion with multiple prompt types nor integrate user-specified prompts with visible appearance constraints, which is the focus of our work.

3. Proposed Method

3.1. Preliminary: Conditional Latent Diffusion

DMs synthesize images by progressively denoising a sample that is initially drawn from random noise. Since performing this process directly in pixel space is computationally expensive and scales with image resolution, recent approaches adopt latent diffusion models, which operate in the latent space of a VAE. In this setting, the DM incorporates text embeddings and ControlNet-based conditioning signals into the denoising process.

For an input image I, the VAE encoder produces the clean latent $z_0 = \mathcal{E}(I)$. The noisy latent z_t at timestep t is obtained via the forward diffusion process:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon, \tag{1}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$ is defined by the diffusion schedule, and $\epsilon \sim \mathcal{N}(0, I)$.

To synthesize images aligned with the given conditions from z_t, the denoising U-Net ϵ_θ and the ControlNet are trained with the following objective:

$$\mathcal{L} = \mathbb{E}_{z_0, \epsilon, t, c_{\text{te}}, c_{\text{pr}}} \left[\left\| \epsilon - \epsilon_\theta(z_t, t, c_{\text{te}}, c_{\text{pr}}) \right\|_2^2 \right], \tag{2}$$

where c_{te} denotes the text embedding extracted from a CLIP encoder [44], and c_{pr} denotes the ControlNet-derived conditioning signal produced from the user-specified prompt.

3.2. Promptable Human Amodal Completion

PHAC aims to synthesize a complete person image from an occluded input while preserving the input's visible ap-

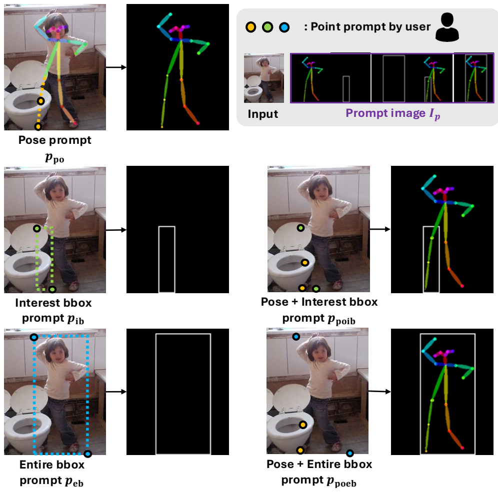

Figure 2. **User prompts**. Users specify the intended pose with point prompts, which we use to condition the model. For the pose prompt p_{po}, we use OpenPose [4] to detect the visible joints, show them to the user, who then adds the missing joints for the desired pose. Alternatively, the user selects two points to specify a bbox prompt, choosing either an interest-region bbox p_{ib} or an entire-region bbox p_{eb}. To provide fine-grained control, the pose and bbox prompts can be combined, yielding p_{poib} or p_{poeb}. To make effective use of the spatial information, we convert the point coordinates into a prompt image I_{p} and use it as a conditioning input.

pearance and aligning with user-specified prompts. Prior amodal completion methods make different input assumptions: they either require an accurate visible-region mask as input [30, 59, 67] or do not accept user-specified prompts [40, 53]. Although PGPIS takes as input a source image and a target pose, which are similar to our inputs, it is restricted to a single prompt type (e.g., pose map [3, 31] or UV map [12, 28]). Our approach takes a user prompt as a conditioning input, including a pose prompt and a bbox prompt that specifies the region to synthesize. For a pose prompt, given the visible joints, the user only needs to add the additional joints for the desired pose; for a bbox prompt, the box can either specify only the interest region for synthesis or span the entire region, including visible regions. The pose and bbox prompts can also be combined. As illustrated in Fig. 2, the user selects any of five prompt types and uses it as the user prompt P, defined as:

$$P \in \{p_{\text{po}}, p_{\text{ib}}, p_{\text{eb}}, p_{\text{poib}}, p_{\text{poeb}}\}, \tag{3}$$

where p_{po} is the pose prompt; p_{ib} and p_{eb} are the bbox prompts for the interest region and the entire region, respectively. p_{poib} and p_{poeb} denote composite prompts that combine a pose prompt with an interest-region bbox or an entire-region bbox, respectively.

The PHAC network Φ_{PHAC} takes an incomplete human image I_{ic} and the user prompt P as inputs and outputs a complete human image I_{c} as follows:

$$I_{\text{c}} = \Phi_{\text{PHAC}}(I_{\text{ic}}, P). \tag{4}$$

The proposed method injects the prompt-based conditioning signal c_{pr}, which is computed by a ControlNet Φ_{CN} specialized for each prompt type, into the coarse image generation network Φ_{CIG}. Conditioned on c_{pr}, Φ_{CIG} takes the incomplete image I_{ic} and the visible mask M_{v} as inputs and produces a coarse complete image I_{cc}, formulated as:

$$c_{\text{pr}} = \Phi_{\text{CN}}(P),\ I_{\text{cc}} = \Phi_{\text{CIG}}(I_{\text{ic}}, M_{\text{v}}, c_{\text{pr}}). \tag{5}$$

To mitigate degradation of the visible appearance arising from latent space denoising and VAE reconstruction, the refinement network Φ_{RF} takes as input the incomplete image I_{ic}, the coarse completion I_{cc}, the visible-region mask M_{v}, and the invisible-region mask M_{iv}, and outputs the refined complete image I_{rc} as:

$$I_{\text{rc}} = \Phi_{\text{RF}}(I_{\text{ic}}, I_{\text{cc}}, M_{\text{v}}, M_{\text{iv}}). \tag{6}$$

The coarse image generation network and the refinement network are described in detail in Secs. 3.3 and 3.4, respectively.

3.3. Coarse Image Generation Network

User prompt. In PHAC, the user prompt encodes spatial information about the desired pose, and the goal is to synthesize an image conditioned on this information. Prompts that impose strong constraints on occluded regions, such as 3D cues (e.g., depth or normal maps) or an invisible mask, can serve as high-quality prompts. In practice, however, such prompts are difficult for users to create and provide to the network. To balance prompt effectiveness with usability, we instead define user-friendly point prompts that require only a small set of points on the image. To obtain the pose prompt p_{po}, we first detect the visible joints with an off-the-shelf 2D pose estimator [4]. Given the detected visible joints, the user adds the missing joints for the desired pose; these user-specified points form the pose prompt p_{po} (orange points in Fig. 2). For the bbox prompts p_{ib} and p_{eb}, the user specifies two points to define either a bbox over the interest region (light-green points) or a bbox covering the entire region, including visible areas (sky-blue points). To provide a stronger constraint, the user can select a composite prompt that pairs a pose with a bbox, such as p_{poib} or p_{poeb}.

Coarse image generation. To effectively exploit the spatial information in the user prompt P, we convert it into a prompt image I_{p} rather than using raw coordinate values.

The prompt image I_{p} is fed to a prompt-specific ControlNet Φ_{CN} to produce the prompt-conditioning signal c_{pr}, computed as follows:

$$c_{\text{pr}} = \Phi_{\text{CN}}(I_{\text{p}}). \tag{7}$$

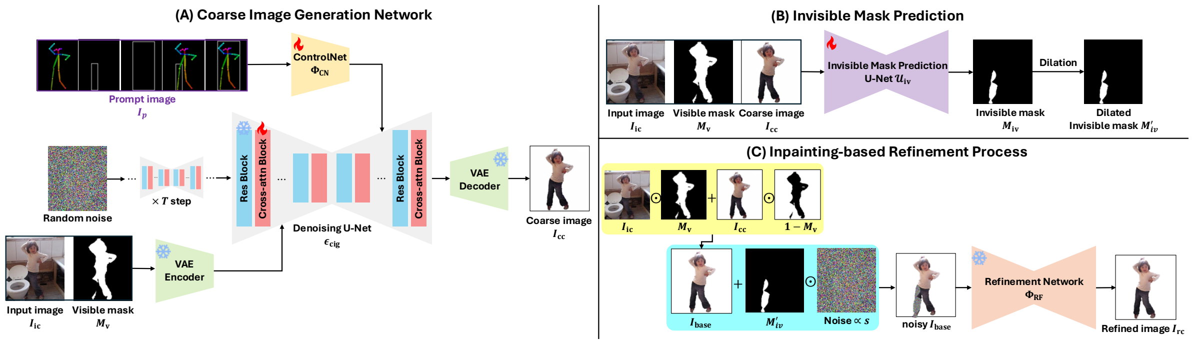

Figure 3. **Method overview**. Given an incomplete image I_{ic} and a user prompt P, our PHAC framework processes them through (A) coarse image generation and (B, C) a refinement stage. In (A), the denoising U-Net ϵ_{cig} starts from random noise and denoises it for T steps to generate a coarse complete image I_{cc}, conditioned on a prompt image I_{p} (see Fig. 2); I_{p} is fed to a prompt-specific ControlNet Φ_{CN} to provide conditioning, and only the cross-attention blocks of ϵ_{cig} are fine-tuned to preserve the pre-trained prior. In (B), invisible mask prediction U-Net $\mathcal{U}_{\text{iv}}$ predicts an invisible mask M_{iv}, which is then dilated to M'_{iv}. In (C), we construct the base composite I_{base} and add low-magnitude noise to the invisible region only. The refinement network Φ_{RF} then takes the noisy I_{base} as input and outputs the refined completion I_{rc}, preserving the visible region while refining the coarse completion and mitigating boundary artifacts.

The text condition c_{te} is obtained from the input image using the CLIP encoder [44]. This step is omitted in Fig. 3 for simplicity. Unlike previous amodal completion methods [30, 59, 67] that used ground-truth (GT) masks, we predict the visible mask M_{v} from the input image using SAM [23] as follows:

$$c_{\text{te}} = \text{CLIP}(I_{\text{ic}}),\ M_{\text{v}} = \text{SAM}(I_{\text{ic}}). \tag{8}$$

Given the incomplete image I_{ic}, the visible mask M_{v}, the text condition c_{te}, and the prompt condition c_{pr}, the denoising U-Net ϵ_{cig} iteratively denoises a latent initialized from random noise to synthesize the coarse completion I_{cc}. The denoising U-Net ϵ_{cig} and the ControlNet Φ_{CN} are jointly trained to minimize the following objective:

$$\mathcal{L} = \mathbb{E}_{z_0,\epsilon,t,c_{\text{te}},c_{\text{pr}}}\left[\|\epsilon - \epsilon_{\text{cig}}(z_t, t, c_{\text{te}}, c_{\text{pr}})\|_2^2\right], \tag{9}$$

where t denotes the diffusion timestep with $0 \le t < 1000$. The clean latent z_0 from the input image I_{ic} and the visible mask M_{v} is obtained by the VAE encoder $\mathcal{E}$ as:

$$z_0 = \mathcal{E}(I_{\text{ic}}) \oplus \mathcal{E}(M_{\text{v}}), \tag{10}$$

where $\oplus$ denotes the concatenation operation.

To better align the generated image with the ControlNet-derived prompt condition c_{pr}, we fine-tune only the cross-attention blocks of the denoising U-Net ϵ_{cig}, keeping the remaining weights frozen, as in [13]. This enables ϵ_{cig} to generate images that align more effectively with the user prompt P. The remaining blocks are kept frozen to preserve the pre-trained DM prior, yielding plausible human appearance.

3.4. Refinement Network

The coarse completion I_{cc} generated by ϵ_{cig} exhibits degradation of the visible appearance arising from latent space denoising and subsequent VAE reconstruction. A naïve approach is to construct a baseline composite I_{base} by copying pixels from the input image I_{ic} in visible regions and from I_{cc} in invisible regions, guided by the visible mask M_{v}:

$$I_{\text{base}} = I_{\text{ic}} \odot M_{\text{v}} + I_{\text{cc}} \odot (1 - M_{\text{v}}), \tag{11}$$

where $\odot$ denotes element-wise multiplication.

While this approach preserves the visible appearance, it often introduces boundary artifacts at mask boundaries and yields an unnatural composite. To address this, we propose an inpainting-based refinement network that keeps the unmasked regions unchanged and performs a few denoising steps in the masked area to reduce boundary artifacts and produce smooth transitions at mask boundaries.

Invisible mask prediction. To apply the inpainting-based approach, we require an invisible-region mask M_{iv}. The masked region specifies where the output should follow the coarse completion I_{cc}, whereas the unmasked region specifies where information from the incomplete input I_{ic} should be preserved. Because a coarse complete image I_{cc} is available, we predict the invisible mask using a lightweight U-Net $\mathcal{U}_{\text{iv}}$. Given the incomplete image I_{ic}, the coarse completion image I_{cc}, and the visible mask M_{v}, U-Net $\mathcal{U}_{\text{iv}}$ outputs the invisible mask M_{iv} as:

$$M_{\text{iv}} = \mathcal{U}_{\text{iv}}(I_{\text{ic}}, I_{\text{cc}}, M_{\text{v}}). \tag{12}$$

We train the U-Net $\mathcal{U}_{\text{iv}}$ with a loss function defined as a weighted sum of binary cross-entropy (BCE) and Dice losses [35] as follows:

$$\mathcal{L} = \mathcal{L}_{\text{BCE}}(M_{\text{iv}}, M^*_{\text{iv}}) + \lambda_{\text{dice}}\mathcal{L}_{\text{Dice}}(M_{\text{iv}}, M^*_{\text{iv}}), \tag{13}$$

where M^*_{iv} is the GT invisible mask.

Directly applying M_{iv} can exclude boundary pixels and introduce boundary artifacts. We therefore use a dilated

Method	User Prompt	OccThuman2.0 test dataset (synthetic)						AHP test dataset (real)					
		LPIPS* ↓	SSIM ↑	KID* ↓	MSE* ↓	PSNR ↑	Joint Err. ↓	LPIPS* ↓	SSIM ↑	KID* ↓	MSE* ↓	PSNR ↑	Joint Err. ↓
PIDM [3]	2D pose map	126.33	0.797	56.91	27.34	16.80	113.72	143.15	0.804	23.71	25.33	16.61	40.47
MCLD [28]	UV map	115.90	0.833	41.11	18.94	18.37	53.38	121.80	0.853	21.90	14.85	18.97	11.53
pix2gestalt [40]	-	90.11	0.911	16.51	7.58	22.63	36.65	75.73	0.942	5.98	5.22	24.06	10.96
pix2gestalt [40]†	2D pose map	88.58	0.914	16.75	6.93	22.94	31.37	75.25	0.943	6.35	4.87	24.25	10.42
SDHDO [39]	2D pose map	81.39	0.924	16.41	7.05	23.80	43.49	64.19	0.956	6.05	6.05	24.45	9.24
Ours	2D pose map	**49.47**	**0.948**	**6.12**	**4.37**	**25.86**	**23.33**	**38.77**	**0.970**	**1.25**	**3.41**	**26.93**	**6.37**
Ours w/o Refinement	2D pose map	58.99	0.941	7.44	5.06	24.67	23.63	64.40	0.957	2.64	4.54	24.54	7.28

Table 1. **Quantitative comparison for promptable human amodal completion.** We use the same 2D pose map as the user prompt, whereas MCLD uses UV maps following its experimental setting. The best performance for each metric is highlighted in **bold**, while the second-best is underlined. † indicates injection of the same user prompt via a pre-trained ControlNet for fair comparison. * indicates that values are reported $\times 10^3$ for readability.

mask M'_{iv} during the refinement process, computed as:

$$M'_{\text{iv}} = \text{Dilate}(M_{\text{iv}}). \tag{14}$$

Inpainting-based refinement process. Rather than synthesizing entirely new RGB values inside the mask as in conventional inpainting, we aim to preserve the coarse completion and visible appearance while reducing boundary artifacts. To this end, we inject a small amount of noise into the masked region, including the occluded area and its boundary, and run a few denoising steps to refine the completion. With this strategy, we keep the unmasked region unchanged, thereby preserving its appearance, and add only low-magnitude noise to the masked region. This minimizes damage to the coarse completion and effectively removes boundary artifacts. The refinement network is plug-and-play and can be applied to other DM-based methods to improve performance.

The refinement network Φ_{RF} takes as input the baseline composite I_{base} from Eq. (11) and the dilated invisible mask M'_{iv} from Eq. (14), and outputs the refined complete image I_{rc} as follows:

$$I_{\text{rc}} = \Phi_{\text{RF}}(I_{\text{base}}, M'_{\text{iv}}, s), \tag{15}$$

where s controls the magnitude of noise added to the masked region (noise magnitude $\propto s$). Larger values of s inject more noise, pushing the effective input toward random noise, whereas values of s near zero keep the input close to a clean image.

4. Experimental Results

4.1. Datasets

To train the ControlNet, the coarse image generation DM, and the invisible mask prediction U-Net, we constructed a synthetic dataset called OccThuman2.0 based on THuman2.0 [57]. Specifically, we rendered complete person images from 526 3D human meshes, each from 10 viewpoints. Because the renders contain only upright, centered subjects, we applied flipping, rotation, and scaling augmentations to increase diversity and used the augmented renders as GT complete images. To increase realism, we sampled backgrounds from Places2 [66] and composited random object instances from MSCOCO [27] in front of the subjects to simulate occlusions, yielding 5,260 occluded person images paired with corresponding GT completions. We sampled occlusion ratios from a Gaussian distribution, as in [39]. We used images from 522 subjects for training and images from the remaining 4 subjects for testing. To assess generalization to in-the-wild images, we also evaluated on the AHP real test set [67], which comprises 56 images with diverse occlusion scenarios. Additional details of the OccThuman2.0 dataset construction are provided in the supplementary material.

4.2. Implementation Details

We initialize the coarse image generation DM with the weights of the DMs used in prior amodal completion [40] and HAC [39], and then fine-tune the model. Following [62], we train the ControlNet by creating a trainable copy of the DM encoder. For ControlNet, when using a single prompt type (pose only or bbox only, p_{po}, p_{ib}, p_{eb}), the prompt input is represented with 3 channels. When using pose and bbox jointly (p_{poib}, p_{poeb}), the prompt input contains 6 channels obtained by concatenating the two 3-channel prompts. During training of the DM and ControlNet, we use a constant warm-up schedule [13]. The learning rates are set to 5×10^{-6} for the DM and 5×10^{-5} for ControlNet, with a batch size of 14. The conditioning scale and classifier-free guidance scale are fixed at 1.0 and 2.0, respectively. We apply a dropout rate of 0.05 to the image-conditioning path. In each epoch, we randomly sample one of the 10 views for each subject, yielding 522 images. We train for 1,750 epochs on four RTX A6000 GPUs, with a total training time of approximately 16 hours.

The invisible mask prediction U-Net $\mathcal{U}_{\text{iv}}$ takes as input the coarse completion I_{cc} produced by the coarse image generation DM. Following a stochastic inference scheme [8,

Figure 4. **Qualitative comparison on the AHP test dataset.** (A) Partial RGB results; (B) Generated images with different seeds. PGPIS baselines (PIDM, MCLD) frequently hallucinate training set appearances. Amodal completion baselines (pix2gestalt, SDHDO) do not preserve the visible appearance and often violate the pose condition. In contrast, our approach yields consistent pose alignment across seeds and preserves the visible regions.

13], the DM generates $N = 16$ outputs per GT image, and during training $\mathcal{U}_{\text{iv}}$ randomly samples one of these outputs for supervision. As with the coarse image generation DM, we use 522 images per epoch and train for 40 epochs. We fix the Dice-loss weight at $\lambda_{\text{dice}} = 0.5$. Training runs on a single RTX 3090 and takes approximately 30 minutes.

4.3. Evaluation Metrics

To assess the appearance of the generated complete human images, we evaluate the quality of reconstructed images using Learned Perceptual Image Patch Similarity (LPIPS), Structural Similarity (SSIM), Kernel Inception Distance (KID), Mean Squared Error (MSE), and Peak Signal-to-Noise Ratio (PSNR). Among these, LPIPS, SSIM, and KID evaluate perceptual quality and visual realism, whereas MSE and PSNR focus on pixel-level reconstruction accuracy. For alignment with user prompts, we report a pose metric given by the pixel-wise joint error between poses predicted on the generated images and those predicted on the GT images, which reflects prompt alignment quality.

4.4. Comparison with Existing Methods

We compare our method with the PGPIS baselines PIDM [3] and MCLD [28], and the amodal completion methods pix2gestalt [40] and SDHDO [39]. For the PGPIS baselines, we provided user prompts in the format that each model was trained to accept and supplied masked images as input. Because pix2gestalt does not accept user prompts, we additionally evaluate pix2gestalt†, which injects the same user prompt via a pre-trained ControlNet for a fair comparison. SDHDO is an HAC method that uses a 2D pose prior. Accordingly, we replace its predicted pose with the user-specified pose used for the other methods.

Table 1 summarizes quantitative results for the PHAC task. On both the synthetic OccThuman2.0 and the real AHP test set, our method outperforms prior approaches on all metrics. PGPIS methods (PIDM, MCLD), which do not adequately preserve the visible appearance, underperform both the amodal completion baselines (pix2gestalt, SDHDO) and our approach. In particular, despite leveraging a richer UV map prompt, MCLD tends to overfit to the training data and performs poorly on both perceptual and reconstruction metrics. For pix2gestalt, adding a 2D pose map condition yields a small gain on OccThuman2.0 but little change on AHP. In contrast, our method performs well on both synthetic and real data and achieves consistent improvements across all metrics.

Qualitative results in Fig. 4 further demonstrate the effectiveness of our method on PHAC. PIDM and MCLD fail to preserve the visible appearance, instead synthesizing images whose appearance reflects dataset-specific biases learned from DeepFashion [29]. PIDM exhibits severe hallucinations; for example, an older man is transformed into a white woman, and the method frequently fails to satisfy the pose condition. Although MCLD aligns better with the target pose than PIDM, it still produces strong appearance hallucinations. As shown in the bottom row of Fig. 4(B), an erroneous UV map causes the model to generate a person wearing the occluder's green clothing instead of the original black garment.

Unlike the PGPIS baselines, pix2gestalt and SDHDO avoid hallucinations in the visible regions and generate plausible person images. However, as in the bottom row of Fig. 4(B), they often over-generate arms or legs irrespective of the specified pose, producing physically implausible results. As shown in Fig. 4(A), they also degrade the visible facial appearance and fail to produce meaningful detail, leading to blurry outputs. Across multiple random seeds, their results are pose-inconsistent and frequently misaligned with the pose condition. In contrast, our method successfully reconstructs the visible region and produces high-quality, physically plausible completions for occluded

User Prompt	LPIPS* ↓	SSIM ↑	MSE* ↓	PSNR ↑	Joint Err. ↓
Pose (p_{po})	49.47	**0.948**	**4.37**	**25.86**	23.33
Interest bbox (p_{ib})	51.83	0.942	5.69	24.99	24.01
Entire bbox (p_{eb})	52.28	0.941	5.66	25.07	28.23
Pose & Interest bbox (p_{poib})	**49.35**	<u>0.947</u>	<u>4.56</u>	<u>25.69</u>	<u>22.15</u>
Pose & Entire bbox (p_{poeb})	<u>49.42</u>	0.946	4.89	25.49	**21.96**

Table 2. **Quantitative comparison across user prompts.** On OccThuman2.0, the pose prompt p_{po} yields the best reconstruction metrics, while bbox prompts ($p_{\text{ib}}, p_{\text{eb}}$) perform worse and exhibit higher joint errors. Combining pose and bbox ($p_{\text{poib}}, p_{\text{poeb}}$) consistently reduces joint error, improving alignment.

s	OccThuman2.0 test dataset				AHP test dataset			
	LPIPS* ↓	SSIM ↑	MSE* ↓	PSNR ↑	LPIPS* ↓	SSIM ↑	MSE* ↓	PSNR ↑
0.1	**49.39**	**0.948**	4.41	25.82	39.02	**0.970**	3.46	26.84
0.3	49.49	**0.948**	<u>4.39</u>	<u>25.83</u>	<u>38.97</u>	**0.970**	<u>3.44</u>	<u>26.88</u>
0.5	<u>49.47</u>	**0.948**	**4.36**	**25.85**	**38.77**	**0.970**	**3.41**	**26.93**
0.7	50.18	0.947	4.43	25.27	39.03	0.969	3.45	26.81
0.9	51.91	0.943	4.77	25.39	40.02	0.967	3.68	26.33

Table 3. **Effect of the refinement noise strength s.** Small s yields limited improvement due to minimal denoising, whereas large s degrades the original appearance.

areas, while maintaining consistent alignment with the pose condition across seeds. All results are obtained via random sampling, not cherry-picking. Additional qualitative results are provided in the supplementary material.

4.5. Ablation Experiments

User prompts. Table 2 presents PHAC results on the OccThuman2.0 dataset with different user prompts. Overall, the pose prompt p_{po} provides the most effective single-prompt guidance, whereas bbox-only prompts ($p_{\text{ib}}, p_{\text{eb}}$) consistently underperform and exhibit higher joint errors. This is likely because bbox prompts mainly constrain the spatial extent of the synthesis region while leaving body configuration ambiguous, allowing diverse poses within the specified box. Combining pose and bbox ($p_{\text{poib}}, p_{\text{poeb}}$) maintains similar perceptual quality to the pose-only setting, while slightly sacrificing reconstruction metrics, but achieves a small and consistent reduction in joint error, indicating improved spatial alignment. Additional analyses are provided in the supplementary material.

Magnitude of noise. Table 3 reports an ablation of the noise strength s used during refinement, where noise is added inside the invisible region to suppress boundary artifacts. We adopt the standard definition $s \in [0, 1]$ (noise magnitude $\propto s$): small s leaves the masked region nearly clean for Φ_{RF}, whereas large s makes it approach random noise. Insufficient noise fails to resolve boundary artifacts, while excessive noise degrades visible regions and deviates from the coarse appearance. Because refinement has minimal effect on pose alignment, we report only image-quality metrics. On both OccThuman2.0 and AHP, $s = 0.5$ yields the best reconstruction and perceptual scores. At $s = 0.5$, we run roughly 40% of the standard denoising steps [42], preserving the visible region and harmonizing the coarse and visible appearances.

Method	LPIPS* ↓	KID* ↓	MSE* ↓	PSNR ↑	Joint Err. ↓
MCLD [28]	59.56 (49%)	11.77 (71%)	7.50 (60%)	23.36 (27%)	31.41 (41%)
pix2gestalt [40]	<u>57.03</u> (37%)	<u>8.14</u> (51%)	<u>6.21</u> (18%)	<u>24.39</u> (27%)	<u>30.65</u> (16%)
SDHDO [39]	59.41 (27%)	10.92 (34%)	6.80 (4%)	24.12 (1%)	33.97 (22%)
Ours	**49.47** (16%)	**6.12** (18%)	**4.37** (14%)	**25.86** (5%)	**23.33** (1%)

Table 4. **Plug-and-play application of the refinement network.** Blue marks results improved by refinement, and % reports the relative improvement over the baseline.

Refinement network. With an appropriate choice of s, the proposed refinement network can be applied in a plug-and-play manner to improve the outputs of other generative methods. Table 4 presents results after applying our refinement network to existing PGPIS and amodal completion baselines. The refinement improves not only perceptual quality and reconstruction metrics but also pose-related measures. In particular, for MCLD [28], which exhibited larger errors, our refinement reduces MSE and KID by approximately 60–70%. Baseline (pre-refinement) scores for the existing methods are reported in Table 1, and the last row isolates the effect of our refinement network on the proposed method. Across both datasets, refinement improves all metrics. Moreover, even before refinement, our method already outperforms prior methods.

5. Conclusion

We introduced a new task, *promptable human amodal completion (PHAC)*, which completes occluded human images while satisfying both visible appearance constraints and user-specified prompts. Our framework injects point-based pose and bbox prompts via ControlNet modules and fine-tunes only the cross-attention blocks, achieving strong prompt alignment. Additionally, we propose an inpainting-based refinement module that leaves the visible regions unchanged, applies a few denoising steps to the completed areas, and maintains boundary continuity. Because it operates on generic inputs, the module can be used as a plug-and-play component with existing diffusion-based completion methods. Experiments on amodal completion and PGPIS benchmarks show that our approach achieves better prompt alignment and produces more physically plausible, higher-quality completions than prior work.

Acknowledgement. This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2023-00219700, Development of FACS-compatible Facial Expression Style Transfer Technology for Digital Human).

References

[1] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM Transactions on Graphics (TOG)*, 42 (4):1–11, 2023. 2

[2] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2745–2754, 2017. 3

[3] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5968–5976, 2023. 2, 3, 4, 6, 7

[4] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 4

[5] Helisa Dhamo, Nassir Navab, and Federico Tombari. Object-driven multi-layer scene decomposition from a single image. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5369–5378, 2019. 3

[6] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 2018. 3

[7] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6144–6153, 2018. 2, 3

[8] Guénolé Fiche, Simon Leglaive, Xavier Alameda-Pineda, and Francesc Moreno-Noguer. Mega: Masked generative autoencoder for human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5366–5378, 2025. 6

[9] Patrick Follmann, Rebecca König, Philipp Härtinger, Michael Klostermann, and Tobias Böttger. 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. 3

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2672–2680, 2014. 3

[11] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7297–7306, 2018. 2

[12] Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with pose-constrained latent diffusion. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22768–22777, 2023. 2, 3, 4

[13] Hsuan-I Ho, Jie Song, and Otmar Hilliges. Sith: Single-view textured human reconstruction with image-conditioned diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 538–549, 2024. 2, 5, 6, 7

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020. 2, 3

[15] Cheng-Yen Hsieh, Kaihua Chen, Achal Dave, Tarasha Khurana, and Deva Ramanan. Tao-amodal: A benchmark for tracking any object amodally. *arXiv preprint arXiv:2312.12433*, 2023. 3

[16] Yuan-Ting Hu, Hong-Shuo Chen, Kexin Hui, Jia-Bin Huang, and Alexander G Schwing. Sail-vos: Semantic amodal instance level video object segmentation-a synthetic dataset and baselines. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3105–3115, 2019. 3

[17] Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16860–16870, 2025. 3

[18] 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 (CVPR)*, pages 1125–1134, 2017. 3

[19] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: Interactive video generation via masked-diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8079–8088, 2024. 3

[20] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *IEEE International Conference on Computer Vision (ICCV)*, pages 127–135, 2015. 3

[21] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4019–4028, 2021. 3

[22] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 3, 5

[24] Jaerin Lee, Daniel Sungho Jung, Kanggeon Lee, and Kyoung Mu Lee. Semanticdraw: towards real-time interactive content creation from image diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13021–13030, 2025. 3

[25] Ke Li and Jitendra Malik. Amodal instance segmentation. In *European Conference on Computer Vision (ECCV)*, pages 677–693. Springer, 2016. 3

[26] Zhixuan Li, Weining Ye, Juan Terven, Zachary Bennett, Ying Zheng, Tingting Jiang, and Tiejun Huang. Muva: A new large-scale benchmark for multi-view amodal instance segmentation in the shopping scenario. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 23504–23513, 2023. 3

[27] 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 (ECCV)*, pages 740–755. Springer, 2014. 6

[28] Jiaqi Liu, Jichao Zhang, Paolo Rota, and Nicu Sebe. Multi-focal conditioned latent diffusion for person image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16019–16028, 2025. 1, 2, 3, 4, 6, 7, 8

[29] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1096–1104, 2016. 3, 7

[30] Zhengzhe Liu, Qing Liu, Chirui Chang, Jianming Zhang, Daniil Pakhomov, Haitian Zheng, Zhe Lin, Daniel Cohen-Or, and Chi-Wing Fu. Object-level scene deocclusion. In *ACM SIGGRAPH 2024 Conference*, pages 1–11, 2024. 2, 4, 5

[31] Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6420–6429, 2024. 2, 3, 4

[32] Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10806–10815, 2021. 3

[33] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 2017. 3

[34] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3d interacting hand pose estimation by hand de-occlusion and removal. In *European Conference on Computer Vision (ECCV)*, pages 380–397. Springer, 2022. 2, 3

[35] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *International Conference on 3D vision (3DV)*, pages 565–571. Ieee, 2016. 5

[36] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 3

[37] Khoi Nguyen and Sinisa Todorovic. A weakly supervised amodal segmenter with boundary uncertainty estimation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7396–7405, 2021. 3

[38] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 3

[39] Seung Young Noh and Ju Yong Chang. Stable diffusion-based approach for human de-occlusion. In *ACM International Conference on Multimedia*, pages 10044–10053, 2025. 1, 2, 3, 6, 7, 8

[40] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt: Amodal segmentation by synthesizing wholes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3931–3940. IEEE Computer Society, 2024. 2, 3, 4, 6, 7, 8

[41] Dim P Papadopoulos, Youssef Tamaazousti, Ferda Ofli, Ingmar Weber, and Antonio Torralba. How to make a pizza: Learning a compositional layer-based gan model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8002–8011, 2019. 2, 3

[42] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 8

[43] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3014–3023, 2019. 3

[44] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PmLR, 2021. 3, 5

[45] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning (ICML)*, pages 8821–8831. Pmlr, 2021. 3

[46] N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt: Watch and learn 2d amodal representation from time-lapse imagery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9356–9366, 2022. 3

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2, 3

[48] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)*, 35:36479–36494, 2022. 3

[49] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3408–3416, 2018. 3

[50] Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. *Advances in Neural Information Processing Systems (NeurIPS)*, 28, 2015. 3

[51] Khiem Vuong, N Dinesh Reddy, Robert Tamburo, and Srinivasa G Narasimhan. Walt3d: Generating realistic training

data from time-lapse imagery for reconstructing dynamic objects under occlusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9514–9524, 2024. 3

[52] Yufu Wang, Yu Sun, Priyanka Patel, Kostas Daniilidis, Michael J Black, and Muhammed Kocabas. Prompthmr: Promptable human mesh recovery. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1148–1159, 2025. 3

[53] Katherine Xu, Lingzhi Zhang, and Jianbo Shi. Amodal completion via progressive mixed context diffusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9099–9109, 2024. 2, 3, 4

[54] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1316–1324, 2018. 3

[55] Xiaosheng Yan, Feigege Wang, Wenxi Liu, Yuanlong Yu, Shengfeng He, and Jia Pan. Visualizing the invisible: Occluded vehicle segmentation and recovery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7618–7627, 2019. 2, 3

[56] Jie Yang, Ailing Zeng, Feng Li, Shilong Liu, Ruimao Zhang, and Lei Zhang. Neural interactive keypoint detection. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15122–15132, 2023. 3

[57] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5746–5756, 2021. 6

[58] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28003–28013, 2024. 3

[59] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3784–3792, 2020. 2, 3, 4, 5

[60] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5907–5915, 2017. 3

[61] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7982–7990, 2021. 3

[62] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 2, 3, 6

[63] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7713–7722, 2022. 3

[64] Ruifei Zhang, Xiangru Lin, Wei Zhang, Jincheng Lu, Xuekuan Wang, Xiao Tan, Yingying Li, Errui Ding, Jingdong Wang, and Guanbin Li. Interactive 3d object detection with prompts. In *European Conference on Computer Vision (ECCV)*, pages 140–157. Springer, 2024. 3

[65] Chuanxia Zheng, Duy-Son Dao, Guoxian Song, Tat-Jen Cham, and Jianfei Cai. Visiting the invisible: Layer-by-layer completed scene decomposition. *International Journal of Computer Vision*, 129:3195–3215, 2021. 3

[66] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1452–1464, 2017. 6

[67] Qiang Zhou, Shiyin Wang, Yitong Wang, Zilong Huang, and Xinggang Wang. Human de-occlusion: Invisible perception and recovery for humans. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3691–3701, 2021. 2, 3, 4, 5, 6

[68] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232, 2017. 3

[69] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1464–1472, 2017. 3

[70] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2347–2356, 2019. 3