

Amodal Depth Anything: Amodal Depth Estimation in the Wild

Zhenyu Li¹, Mykola Lavreniuk², Jian Shi¹, Shariq Farooq Bhat¹, Peter Wonka¹
¹KAUST, ²Space Research Institute NASU-SSAU



Figure 1. **Amodal Depth Estimation in the Wild.** For each image, we present the amodal depth estimation with the target object outlined in black. Our model demonstrates strong generalization across diverse scenes, including both *indoor* and *outdoor* environments for accurate depth estimation for occluded parts of objects. Best viewed in color.

Abstract

Amodal depth estimation aims to predict the depth of occluded (invisible) parts of objects in a scene. This task addresses the question of whether models can effectively perceive the geometry of occluded regions based on visible cues. Prior methods primarily rely on synthetic datasets and focus on metric depth estimation, limiting their generalization to real-world settings due to domain shifts and scalability challenges. In this paper, we propose a novel formulation of amodal depth estimation in the wild, focusing on relative depth prediction to improve model generalization across diverse natural images. We introduce a new large-scale dataset, Amodal Depth In the Wild (ADIW), created using a scalable pipeline that leverages segmentation datasets and compositing techniques. Depth maps are generated using large pre-trained depth models, and a scale-and-shift alignment strategy is employed to refine and blend depth predictions, ensuring consistency in ground-truth annotations. To tackle the amodal depth task, we present

two complementary frameworks: Amodal-DAV2, a deterministic model based on Depth Anything V2, and Amodal-DepthFM, a generative model that integrates conditional flow matching principles. Our proposed frameworks effectively leverage the capabilities of large pre-trained models with minimal modifications to achieve high-quality amodal depth predictions (Fig. 1). Experiments validate our design choices, demonstrating the flexibility of our models in generating diverse, plausible depth structures for occluded regions. Our method achieves a 50.7% improvement in RMSE over the previous SoTA on the ADIW dataset.

1. Introduction

Monocular depth estimation is a foundational task in computer vision and generative modeling [5, 10, 27] as it provides depth perception from a single image without stereo cues. However, while recent methods, e.g. [2, 4, 10, 23, 46], focus solely on estimating depth for visible pixels, humans can intuitively perceive the complete 3D

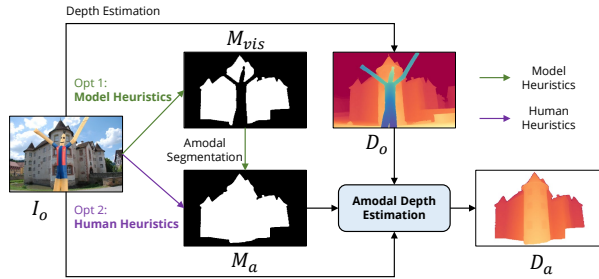


Figure 2. **Amodal Depth Estimation Pipeline.** Given an input image, users can generate the amodal mask for the depth estimator in two ways: **(1) Model Heuristics:** click the target object, apply SAM [21] to generate modal mask, then use amodal segmentation methods to estimate amodal mask, **(2) Human Heuristics:** manually draw the amodal mask. Our model estimates amodal depth based on original observation image I_o , the observed depth map D_o , and the amodal mask M_a . Note that we can get visible and invisible masks from both pipelines. Our model only focuses on the amodal depth (*i.e.*, invisible part), whereas the visible part depth can be directly derived from the observed depth map D_o .

geometry of objects, even when only parts of them are visible. Amodal depth estimation is the task of predicting depth values for the occluded (invisible) parts of objects [17, 38]. It enables accurate 3D reconstruction and novel view synthesis from a single view, significantly reducing dependence on costly multi-view systems or LiDAR [1]. This capability is crucial for advancing AR/VR, robotics, autonomous driving and digital twin technologies [1]. In this paper, we aim to address this under-explored question: *Can models effectively perceive the geometry of the invisible parts of objects in a scene?*

Unlike inpainting methods [11, 39, 50], which reconstruct missing image regions, amodal depth estimation is a novel task that extends traditional amodal segmentation [28, 48, 49] into the depth domain, by predicting the depth of occluded object parts. In this task, given an input image with a target amodal mask, the objective is to infer depth values for the occluded object regions. While amodal segmentation benefits from human annotations to collect training samples [17, 38], there is currently no device capable of collecting ground-truth depth data for occluded parts of objects at scale in real-world scenes.

To address this challenge, prior approaches to amodal depth estimation have relied on synthetic datasets [17, 38]. However, generating these datasets is both time-consuming and hard to scale, often requiring manually placing occluders one by one [17, 38]. Additionally, synthetic data lacks the complexity and diversity of real-world scenes, resulting in a domain gap that limits the generalization of models trained on such data. Furthermore, these previous approaches only consider metric depth, where the goal is to estimate real-world distances for occluded regions. Re-

liance on metric depth, which naturally struggles to generalize to unseen scenarios with limited data [7, 29], exacerbates the models’ poor zero-shot performance on real-world images [6, 34, 46].

To overcome these limitations, we propose a novel formulation of **amodal depth estimation in the wild**, focusing on *relative* depth. Recent advancements in depth estimation models, such as Depth Anything [46], have enabled the generation of high-quality relative depth maps from natural images. Focusing on relative depth allows us to train models using real-world data, leveraging the depth relationships within scenes to achieve better generalization without relying on precise metric measurements.

We introduce a new large-scale real-world dataset, **Amodal Depth In the Wild (ADIW)**, to facilitate the training of models for amodal depth estimation. Our data generation pipeline leverages segmentation datasets to create high-quality datasets for amodal relative depth estimation with scalable annotations. We adopt a compositing approach to create training pairs, where objects are sampled and composited into scenes. Corresponding depth maps are estimated using the large depth foundation model [47]. Given the challenges posed by occlusion in relative depth estimation, we employ a scale-and-shift alignment technique to refine and blend the depth values, ensuring consistency in the final ground-truth depth map used for training.

In this work, we present two complementary frameworks for amodal relative depth estimation, targeting both deterministic and generative model classes. For deterministic models, we propose Amodal-DAV2, an extension of the Depth Anything V2 (DAV2) model [47]. This framework incorporates additional guidance layers to reason about depth in occluded regions. For generative models, we present Amodal-DepthFM, an adaptation of the DepthFM model [14], which uses a conditional flow matching approach to estimate depth in occluded regions. In both frameworks, we leverage the power of large pre-trained models with minimal modifications to effectively guide them toward predicting depth for occluded parts of objects.

Experimental results indicate that our method achieves SoTA amodal depth estimation on the ADIW dataset and demonstrates strong zero-shot performance on real-world images (Fig. 1). Comprehensive ablation studies further validate the effectiveness of our design choices, showing that object-level supervision and guidance signals significantly enhance prediction accuracy. Moreover, we demonstrate the flexibility of our models in generating diverse plausible structures for occluded regions, illustrating their adaptability to different occlusion scenarios. In summary, our contributions are:

- A novel task formulation of amodal depth estimation focusing on relative depth, enabling improved generalization capabilities compared to previous metric-based

- amodal depth estimation.
- ADIW, a large-scale real-world dataset, generated from real-world images using a scalable pipeline with segmentation datasets and compositing techniques, coupled with a scale-and-shift alignment strategy.
- Two novel frameworks for amodal depth estimation, Amodal-DAV2 and Amodal-DepthFM, leveraging large-scale pre-trained models with minimal modifications to achieve high-quality amodal depth predictions.

2. Related Work

2.1. Monocular Depth Estimation

Monocular depth estimation, which predicts depth from a single image, has made substantial progress in recent years [2, 4, 10, 19, 24, 46]. Early methods focused on domain-specific metric depth estimation, assuming that training and test images share similar characteristics [3, 10, 12, 23, 25]. However, this in-domain focus poses challenges for generalization to unseen domains. To address this, recent research has shifted toward cross-domain relative depth estimation, where models infer the relative depth relationships between pixels [4, 6, 34, 46, 47]. For instance, MiDaS [34] aggregates multiple datasets and trains the model in a relative depth setting, achieving superior zero-shot performance. ZoeDepth [4] demonstrates that a strong relative depth model can effectively generalize to metric depth estimation through fine-tuning. Depth Anything [46] adopts a semi-supervised strategy that combines large-scale labeled and unlabeled data to further boost the model performance. Generative approaches have also emerged, with methods like Marigold [19, 22, 53] repurposing denoising UNet models [37] alongside fixed VAE encoders [42] for depth estimation [19], and DepthFM integrates flow matching principles into the depth estimation pipeline [14]. Despite advancements in monocular depth estimation, a major challenge remains: depth estimation in occluded regions. Existing methods focus only on visible areas, leaving unseen parts unaddressed. Our work addresses this gap by enabling direct depth estimation for invisible object regions.

2.2. Amodal Perception

Prior work in amodal perception has primarily focused on tasks such as inpainting [9, 28, 49], segmentation [20, 26, 31, 54], and detection [15, 18]. Recently, amodal depth estimation has emerged as a new challenge in amodal perception [17, 38], which aims to predict the depth of occluded regions of objects. For instance, [38] proposed the Amodal-SynthDrive dataset for synthetic driving scenes and adopted VIP-DeepLab [32] with multi-headed outputs to estimate the depth at different occlusion levels. Similarly, [17] introduced the indoor-focused Amodal-3D-FRONT dataset and an iterative approach leveraging amodal masks to estimate

depth in occluded areas. These works rely heavily on synthetic datasets due to the difficulty of capturing ground-truth depth behind occlusions in real-world settings. This reliance on synthetic data limits model generalization to real-world scenes, and their focus on metric depth increases the difficulty of accurate depth prediction in the zero-shot setting. In contrast, our approach adopts a relative depth-based solution using large-scale pre-trained models and realistic synthetically generated datasets, enabling strong generalization to diverse real-world images.

2.3. Depth Inpainting

Depth inpainting, or depth completion, traditionally focuses on filling in missing depth values for visible regions based on sparse depth inputs [41, 43–45, 52]. Some methods aim to inpaint depth around occlusions to create novel 3D views [11, 16, 39], this differs from amodal depth estimation, which specifically targets object-level occlusions. For example, 3D Photography [39] uses shared edge guidance to combine RGB and depth inpainting, while SLIDE [16] introduces a soft-layering strategy to preserve visual details in novel views. Recently, diffusion models have shown strong image generation capabilities [37]. Invisible Stitch [11] uses a Stable Diffusion model variant [37] to fill in missing image regions, generating novel views [30], followed by a depth painting network to fill holes in the depth map with guidance from the novel-view image. Unlike these methods, which implicitly or explicitly rely on color priors and focus on visual quality, our approach directly estimates occluded region depth, prioritizing depth accuracy independently of color information or visual coherence. This distinction shifts the focus from achieving realistic view synthesis to predicting occluded depths with high geometric accuracy, an approach that enhances scene understanding for real-world applications.

3. Method

3.1. Task Definition

Amodal depth estimation extends the concept of amodal segmentation by predicting depth information for occluded areas. In traditional amodal segmentation [28, 48, 49], given an input image and a segmentation mask of the visible portion of an object, the goal is to predict the complete object mask, including the occluded (invisible) part. Similarly, amodal depth estimation aims to estimate the depth values for the occluded regions, given an input observation image I_o , a corresponding observation depth map D_o of the input image, and a target amodal segmentation mask M_a .

Different from earlier works that formulate amodal depth estimation as a *metric* depth estimation task [17, 38], we propose a novel formulation where the goal is to predict the *relative* depth of the occluded parts of the object for

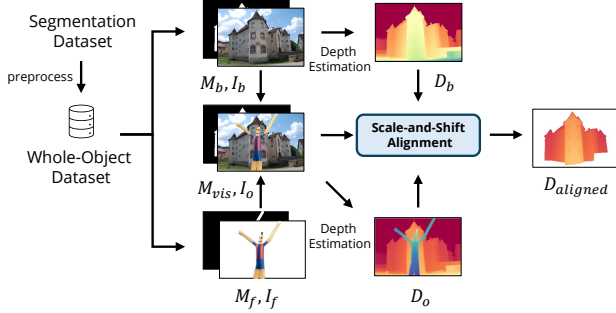


Figure 3. **Constructing Training Data.** We use the method from [28] to convert an initial segmentation dataset into a whole-object dataset. Next, we sample and composite images to create training pairs. Due to occluders, the relative depth predictions differ between the composite and background images, so we apply scale-and-shift alignment for consistent depth blending.

the input image *in the wild*. Recent models, such as Depth Anything [46, 47] generate high-quality relative depth maps from natural images, allowing us to train amodal depth models using real-world data as shown in Sec. 3.2.

3.2. Dataset Collection

We present **Amodal Depth In the Wild (ADIW)**, a large-scale dataset specifically designed for predicting occluded object relative depths in real-world scenes. Creating a natural image dataset at scale for amodal depth estimation is challenging due to the lack of annotations for hidden regions. No existing device can capture ground-truth depth data for occluded parts at scale in real-world settings. Previous efforts have generally relied on synthetic datasets [17, 38], which, although valuable, fall short in capturing the complexity and diversity of real-world scenes. Additionally, these approaches require manually placing occluders, limiting scalability. ADIW overcomes these by generating training data from real-world images.

We create paired data by overlaying objects over natural images following [28], as illustrated in Fig. 3. Our method leverages segmentation datasets, which are more scalable than traditional depth estimation datasets [21, 35, 46]. Specifically, we employ the Segment Anything model [21] to automatically generate segmentation masks from the SA-1B dataset, forming the initial segmentation dataset. Next, we apply the heuristic algorithm from [28] to filter out incomplete objects, creating a whole-object dataset.

In our approach, each image I_b (the background image) contains at least one complete object. We then sample an occluder object I_f (the foreground image) and superimpose it onto I_b , forming the assembled observation image I_o . Both I_o and I_b are then processed through the Depth Anything V2 model [47] (ViT-G) to obtain relative depth maps D_o and D_b , respectively. Importantly, the depth val-

ues for the background object in D_o and D_b differ due to the presence of the foreground object, which changes perceived depth. Both depth maps are scaled to the range $[0, 1]$.

To ensure consistent training labels, we apply the scale-and-shift alignment algorithm [34] to align the background object’s depth values across the two depth maps. The scale factor s and shift factor t are computed as:

$$(s, t) = \operatorname{argmin}_{s, t} \sum_{i=1}^N (s d_i^b + t - d_i^o)^2, \quad i \in M_{vis}, \quad (1)$$

where d_i^o and d_i^b are the depth values of pixels in the visible part of the background object in D_o and D_b , respectively, and N denotes the total number of valid pixels in the visible mask M_{vis} of the background object. The aligned depth map $D_{aligned}$ is then calculated as:

$$D_{aligned} = s D_b + t, \quad (2)$$

serving as the ground-truth map for model training. This procedure generates a dataset of 564K images with amodal depth labels.

As shown by [46], relative depth estimators [6, 34] generalize better than metric depth estimators [2, 23, 25]. Following this insight, we produce relative depth maps for amodal depth estimation in natural scenes. Moreover, our data generation approach is also adaptable for metric amodal depth estimation by omitting the scale-and-shift alignment and utilizing metric depth models [7, 29].

3.3. Amodal Depth Estimator

In this work, we aim to leverage large pre-trained depth models by fine-tuning them specifically for amodal depth estimation. Our goal is to make minimal modifications to the original network architectures, preserving the capabilities of the pre-trained weights while introducing the necessary adjustments to guide the model for amodal depth prediction. By integrating additional guidance into the networks, we enable them to predict depth values for occluded parts of objects.

Depth estimation models fall into two types: deterministic and generative models [13]. In this section, we introduce a dedicated strategy for adapting each model class to the amodal depth estimation task, as illustrated in Fig. 4 and Fig. 5. For the deterministic model, we adopt the Depth Anything V2 (DAV2) [47], a highly representative and top-performing pre-trained model. For the generative model, we select DepthFM [14], known for its superior ability to capture depth details and fast inference speed.

3.3.1. Amodal-DAV2

Model Structure. The Amodal-DAV2 framework introduces subtle structural modifications to the Depth Anything V2 (DAV2) model’s image encoder to retain its pre-trained

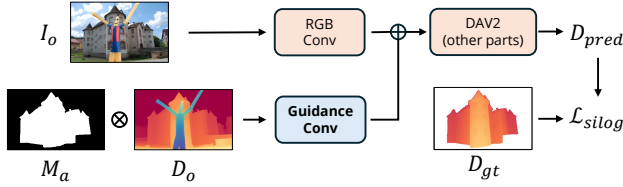


Figure 4. **Amodal-DAV2 Framework Structure.** Amodal-DAV2 modifies the DAV2 image encoder to take additional guidance channels along with RGB.

knowledge while enhancing its capacity in amodal depth estimation. The original DAV2 model employs a Vision Transformer (ViT) architecture [8], which begins with an RGB convolution applied to the input image. As shown in Fig. 4, we add a Guidance Convolution (Conv) [40] layer in parallel with the RGB Conv layer, enabling the encoder to accept additional guidance channels for the observation depth map D_o and amodal mask M_a . To ensure compatibility with the pre-trained model, the kernel weights of the Guidance Conv layer are initialized to zero, allowing the model to initially ignore the additional guidance information. Additionally, we incorporate layer normalization into the input features of the DPT head [33] to stabilize training and improve convergence.

Training method. We fine-tune the entire Amodal-DAV2 framework during training, utilizing the standard scale-invariant loss \mathcal{L}_{si} [10] as our objective function:

$$\mathcal{L}_{si} = \alpha \sqrt{\frac{1}{N} \sum_{i \in M_a} g_i^2 - \frac{\lambda}{N^2} \left(\sum_{i \in M_a} g_i \right)^2},$$

where $g_i = \log \tilde{d}_i - \log d_i$, with \tilde{d}_i and d_i representing the predicted depth and the ground truth depth, respectively. N denotes the number of valid pixels on the amodal mask M_a . Although our primary focus is on the accuracy of depth predictions for occluded (invisible) parts of the objects, we supervise the model using the entire object’s depth. This holistic approach helps the model better understand the overall scene structure, leading to improved performance.

3.3.2. Amodal-DepthFM

Framework and Training. We adapt DepthFM [14] for amodal depth estimation using a similar approach. Given pairs of an image I , its observation depth map D_o , and corresponding amodal mask M_o , we fine-tune a conditional flow matching model to achieve amodal depth estimation. The objective is defined as:

$$\min_{\theta} \mathbb{E}_{t, z, p(x_0)} \|v_{\theta}(t, \phi_t(x_0)) - (x_1 - x_0)\|,$$

where x_1 represents the encoded depth samples in the latent space, and the starting point x_0 corresponds to an encoded representation of the input image. The latent flow

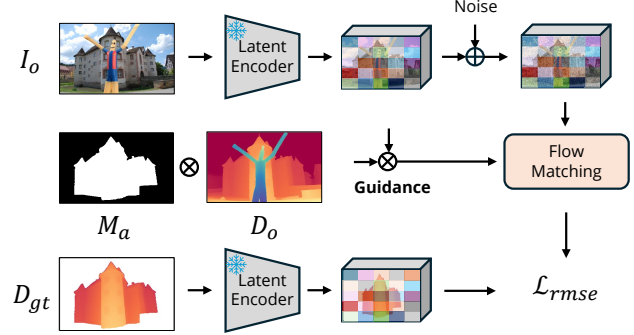


Figure 5. **Amodal-DepthFM Framework Structure.** Amodal-DepthFM modifies the DepthFM denoising UNet encoder to take additional guidance channels along with RGB latent code.

is conditioned on the guidance provided by z , which includes the input image latent code, observation depth map D_o , and amodal mask M_a . Here, t denotes the timestamp, and $\phi_t(x|z) = tx_1 + (1-t)x_0$.

During training, we apply noise augmentation, introducing Gaussian noise $\mathcal{N}(x, \sigma_{min})$ to the data samples. Consequently, the conditional probability path is modeled as $p_t(x|z) = \mathcal{N}(x|tx_1 + (1-t)x_0, \sigma_{min}^2 \mathbf{I})$. For additional details, we refer readers to [14].

In the U-Net architecture used for the flow network v_{θ} , we modify the first input Conv layer to accommodate the additional guidance channels. Specifically, we extend the channel dimensions to accept extra guidance information D_o and M_a alongside the image latent code. To effectively leverage the pre-trained model, we initialize the first Conv layer with a combination of pre-trained weights for the primary eight channels (corresponding to the depth and image latent codes) and zero-initialized weights for the two additional channels dedicated to guidance inputs D_o and M_a .

Scale-and-Shift Alignment Inference. While the Amodal-DAV2 model directly regresses the amodal depth map, the Amodal-DepthFM model operates in the latent space, predicting the amodal depth latent code, which is then decoded into a depth map using a latent decoder. However, learning consistent amodal depth can be challenging for Amodal-DepthFM. To enhance the predicted depth map, we employ the scale-and-shift alignment [34, 36] during inference.

Rather than relying solely on the model’s output, we enhance the prediction by blending it with the observed depth map using a scale-and-shift alignment over the shared visible regions, as described in Eq. 1 and Eq. 2. This technique leverages the information from shared regions between the observation depth map and the amodal prediction. By calculating optimal scale and shift factors, we align the depth values of the occluded regions with the visible areas, thereby enhancing the coherence and consistency of the final depth map.

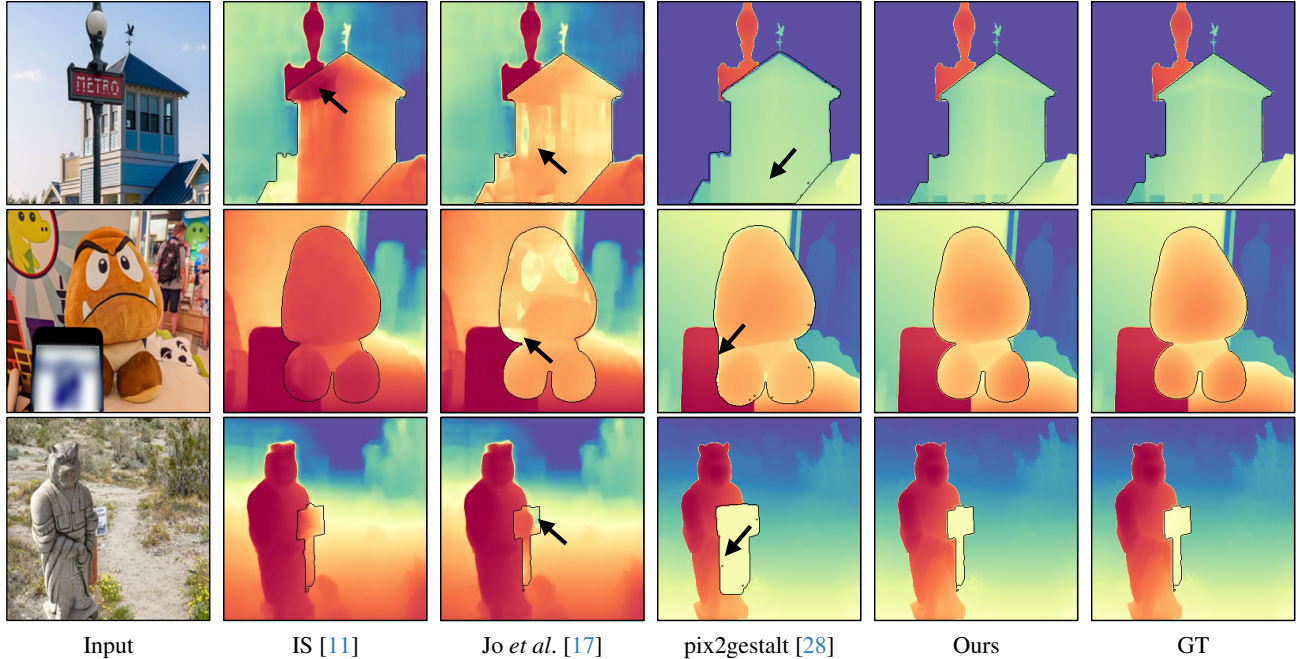


Figure 6. **Qualitative Comparisons on the Validation Set of ADIW.** Since Invisible Stitch (IS) [11] and Jo *et al.* [17] use metric depth estimation models [4], the depth maps are shifted compared with other methods that use relative depth estimators. Only officially released models are taken into account in this qualitative comparison.

Method	Easy		Mid.		Hard		Overall	
	RMSE↓	$\delta(\%)$ ↑	RMSE↓	$\delta(\%)$ ↑	RMSE↓	$\delta(\%)$ ↑	RMSE↓	$\delta(\%)$ ↑
pix2gestalt [‡] [28]	5.067	90.353	4.818	90.271	5.641	84.064	5.114	88.717
Invisible Stitch [‡] [11]	7.584	83.317	7.435	81.359	7.393	80.036	7.481	81.781
Sekkat <i>et al.</i> [†] [38]	11.195	51.052	11.404	51.267	11.150	49.222	11.264	49.222
Sekkat <i>et al.</i> ^{†‡} [38]	5.245	88.705	5.254	88.722	5.044	87.423	5.194	88.367
Jo <i>et al.</i> [†] [17]	10.260	56.118	10.227	55.637	9.982	54.687	10.172	55.545
Jo <i>et al.</i> [‡] [17]	7.551	81.933	6.633	82.733	6.488	82.793	6.939	82.451
Jo <i>et al.</i> ^{†‡} [17]	4.624	89.848	4.777	89.915	4.728	87.256	4.712	89.177
Amodal-DAV2-S	3.574	93.650	3.741	93.666	3.734	92.125	3.682	93.251
Amodal-DAV2-B	3.482	93.871	3.620	94.002	3.640	92.622	3.578	93.590
Amodal-DAV2-L	3.324	94.411	3.460	94.566	3.476	93.309	3.418	94.179
Amodal-DepthFM	5.354	88.427	5.398	89.241	5.498	86.970	5.410	88.353
Amodal-DepthFM [‡]	4.622	92.494	4.500	92.982	4.883	91.048	4.645	92.295

Table 1. **Quantitative Comparisons on ADIW Dataset.** We compare our methods against other possible solutions. Results from Invisible Stitch [11] and Jo *et al.* [17] are not directly comparable, as they use metric depth estimation models [4]. [†]: Models retrained on our dataset for fair comparison. [‡]: Scale-and-shift alignment applied for consistent prediction handling (see 3.3). Note that Amodal-DAV2 does not rely on this alignment approach. Best results are in **bold**.

4. Experiments

4.1. Metrics

We follow the standard evaluation protocol proposed in previous monocular metric depth estimation works [2, 10] to evaluate the effectiveness of our framework. We use the root mean squared error (RMSE), the \log_{10} error, and the

accuracy under the threshold δ . Metrics consider the invisible parts of objects, with the RMSE scaled by a factor of 100 for better illustration. We also categorize objects based on their visible ratios into three difficulty levels: easy (0.75, 1], medium (0.5, 0.75], and hard (0, 0.5]. Metrics are calculated separately for each difficulty level to provide a comprehensive evaluation.

	Method	RMSE↓	log ₁₀ ↓	δ(%)↑
①	w/o D_o, M_a	7.549	8.607	70.607
②	w/o M_a	4.369	4.320	91.193
③	w/ align	3.878	3.659	92.037
④	Ours (Full)	3.682	3.538	93.251
⑤	\mathcal{L}_{silog} for inv. only	3.845	3.751	92.608
⑥	\mathcal{L}_{ssi} + align	4.015	3.948	91.852

Table 2. **Ablation Study for Amodal-DAV2.** We investigate the impact of different guidance signals, supervision strategies, and inference techniques.

4.2. Implementation Details

We split our dataset into a training set of approximately 559K samples and a validation set of 4K samples. The Amodal-DAV2 model is trained with a batch size of 32, a learning rate of $1e^{-5}$, and 50K iterations, using the scale-invariant log (silog) loss with $\lambda=0.85$, following previous works [2, 23, 25]. The Amodal-DepthFM model is trained with a batch size of 128, a learning rate of $3e^{-5}$, and 15K iterations. Both models are initialized with depth-pretrained parameters before fine-tuning for amodal depth estimation. We use the Adam optimizer with exponential learning rate decay. To stabilize training, we apply a max gradient norm clip of 0.01. Data augmentation is minimal, with only horizontal flipping. All experiments are conducted on 4 NVIDIA A100 GPUs. By default, we evaluate model performance using the final checkpoint after training. We use our Amodal-DAV2-L for all visualization results as default.

4.3. Main Results

Comparison with Amodal Depth Methods. We compare our methods with existing amodal depth estimation approaches. As these methods were initially trained on synthetic datasets and relied on metric depth formulations, we retrained them on our dataset and adapted them to relative depth estimation. Results in Tab. 1 and Fig. 6 demonstrate a significant performance gap between the previous SoTA amodal depth model [17] and our methods, even after output alignment with our strategy. Notably, Amodal-DepthFM outperforms prior approaches even without the alignment, while our best model, Amodal-DAV2-L, improves performance by 27.4% in terms of RMSE, setting a new SoTA. Previous methods, designed for amodal metric depth estimation, struggle to generalize across varied natural image depth ranges. In contrast, our approach leverages pre-trained DAV2 [47] and DepthFM [14] models, which benefit from extensive prior knowledge of geometry and color, resulting in more accurate amodal depth predictions.

Comparison with Other Solutions. We also evaluate two alternative solutions: Invisible Stitch [11] and pix2gestalt [28]. (1) For Invisible Stitch, we use SD-XL [30] with a ground-truth amodal mask and image cap-

	Method	RMSE↓	log ₁₀ ↓	δ(%)↑
①	w/o align & D_o, M_a	10.211	12.563	56.271
②	w/o align & M_a	5.283	6.231	83.799
③	w/o align	5.410	5.235	88.353
④	Ours (Full)	4.645	3.553	92.295
⑤	\mathcal{L}_{rmse} for inv. only	5.636	3.911	91.739
⑥	\mathcal{L}_{rmse} for scene	4.608	3.809	91.149

Table 3. **Ablation Study for Amodal-DepthFM.** We explore the effectiveness of various guidance signals, supervision strategies, and the alignment inference for Amodal-DepthFM. Unlike Amodal-DAV2, the alignment strategy significantly improves the performance of Amodal-DepthFM.

tion to guide the inpainting. The inpainted image serves as guidance for the stitcher model [11] to perform depth inpainting, and the predicted depth is aligned with the observation depth map. (2) For pix2gestalt [28], we conduct amodal inpainting followed by depth estimation using DAV2 (ViT-G), aligning the predicted depth with the observation depth. Additional implementation details are in the *supplementary materials*.

As illustrated in Fig. 6 and Fig. 7, the accuracy of amodal depth estimation relies heavily on the quality of RGB inpainting in both methods. Even with the ground-truth amodal mask, SD-XL struggles to accurately inpaint target areas, and the stitcher model fails to predict accurate depth for occluded parts, leaving ghosting artifacts in the predicted depth map. Similarly, pix2gestalt suffers cascade errors due to inaccurate inpainting with uncontrollable amodal shapes. Moreover, since DAV2 is trained on complete natural images, applying it directly to inpainted outputs from pix2gestalt (with single foreground objects on white backgrounds) leads to performance degradation. For example, the window boundary in the first case in Fig. 6 is missing from the predicted depth map. In contrast, our amodal depth models directly regress the depth of invisible parts without relying on RGB information, using only the amodal mask as guidance. These results demonstrate that our approach provides strong geometric priors, which could also serve as a useful condition for inpainting methods [51].

4.4. Ablation Studies and Discussion

Guidance and Supervision Strategies. We conduct ablation experiments on Amodal-DAV2-S and Amodal-DepthFM to assess each framework component’s contribution, with results in Tab. 2 and Tab. 3, respectively. For both models, guidance from the observation depth map D_o and amodal mask M_a is crucial for optimal performance (①, ②). While the primary goal and evaluation protocol focus is estimating depth for occluded parts of objects, both frameworks benefit from object-level supervision that includes visible regions (④, ⑤). Interestingly, the scale-and-shift align-

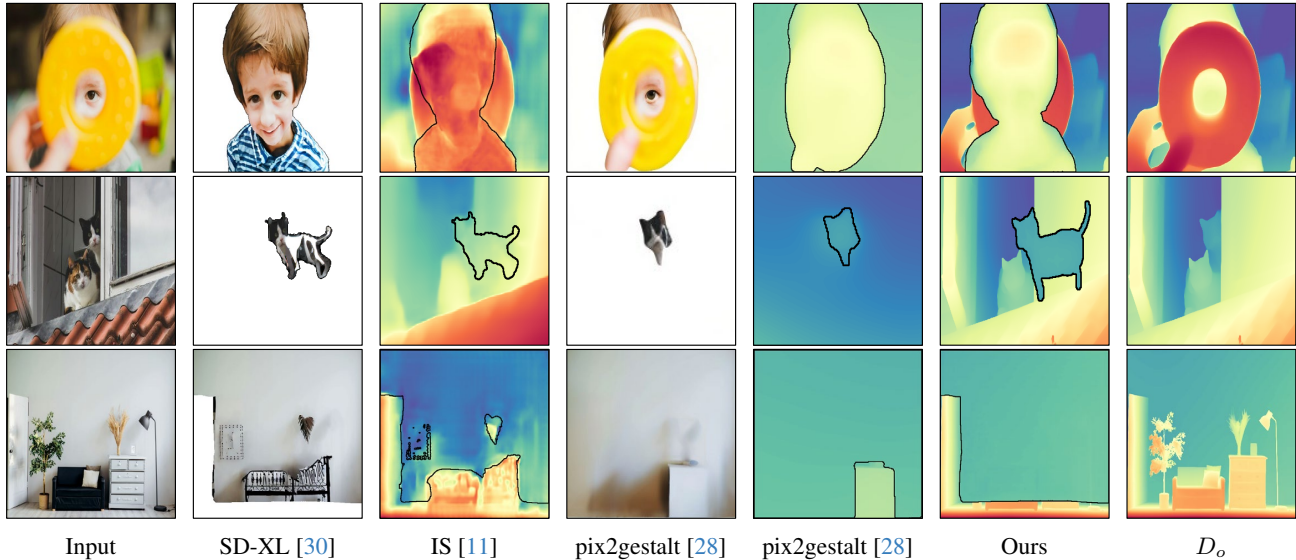


Figure 7. **Qualitative Comparisons on Images in the Wild.** Invisible Stitch (IS) [11] uses SD-XL [30] with a ground-truth visible mask and image caption to inpaint occluded areas, while pix2gestalt [28] completes the occluded areas via amodal inpainting. Both methods suffer from *inaccurate shape completion* and cascade depth errors. The third row highlights the fragile shape completion of previous work: both SD-XL and pix2gestalt fail to remove the foreground cabinets correctly. In the Human Heuristics mode (Fig. 2), our method infers without any RGB inpainting or completion prior and achieves reasonable depth estimations for occluded areas.

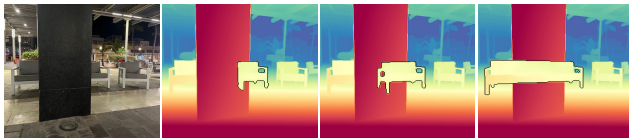


Figure 8. **Results with Different Mask Guidance.** Guided by various amodal masks, our model successfully predicts the corresponding amodal depth maps for the target objects in the image, showcasing flexibility in estimating occluded regions.

ment during inference improves Amodal-DepthFM performance but decreases that of Amodal-DAV2-S, indicating that inconsistent depth estimation may hinder Amodal-DepthFM’s overall performance (③).

Varying Predictions with Different Guidance. Fig. 8 demonstrates the flexibility of our amodal depth estimation model in generating diverse, plausible predictions based on different amodal masks as guidance. This capability is crucial when multiple valid interpretations exist for occluded object parts. For example, the right-hand chair occluded by a wall could vary in width or even be connected to another occluded chair on the left. Our Amodal-DAV2-L adapts the amodal depth output according to the provided mask.

Edge Region v.s., Non-Edge Region. In this experiment, we extract edge areas from the ground-truth depth map and calculate metrics for both edge and non-edge regions in the invisible area. The overall RMSE from Amodal-DAV2-S is 3.682, with the RMSE for the edge region significantly

higher at 7.378, compared to only 3.515 for non-edge regions. This discrepancy highlights the challenges in estimating amodal depth with fine-grained details. This finding aligns with human perception, where it is generally more difficult to accurately recover fine-grained geometry than the coarser structures for occluded areas.

5. Conclusion

We presented a novel approach to amodal depth estimation, focusing on predicting the depth of invisible parts of objects in natural scenes. Our work introduces Amodal Depth In the Wild (ADIW), a large-scale dataset that leverages segmentation datasets and a compositing pipeline, enabling high-quality, real-world amodal depth annotations. We proposed two complementary frameworks for amodal depth estimation: Amodal-DAV2 and Amodal-DepthFM. Amodal-DAV2 leverages the deterministic capabilities of Depth Anything V2 to achieve state-of-the-art performance in relative depth estimation, while Amodal-DepthFM, built on a generative flow matching paradigm, and excels in providing finer details and sharper boundaries in occluded regions. Our experiments highlight that both models benefit from object-level supervision and the importance of guidance signals for improving amodal depth prediction accuracy. This work not only sets a new benchmark for amodal depth estimation but also opens the door for future research in occluded geometry understanding and improving applications such as inpainting and scene reconstruction.

Acknowledgements

This publication is supported by funding from King Abdullah University of Science and Technology (KAUST) - Center of Excellence for Generative AI, under award number 5940.

References

- [1] Jiayang Ao, QiuHong Ke, and Krista A. Ehinger. Image amodal completion: A survey. In *Computer Vision and Image Understanding*. Elsevier, 2023. 2
- [2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018, 2021. 1, 3, 4, 6, 7
- [3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Localbins: Improving depth estimation by learning local distributions. In *European Conference on Computer Vision*, pages 480–496. Springer, 2022. 3
- [4] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 1, 3, 6
- [5] Amlaan Bhoi. Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*, 2019. 1
- [6] Reiner Birkel, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv preprint arXiv:2307.14460*, 2023. 2, 3, 4
- [7] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 2, 4
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 5
- [9] Kiana Ehsani, Roozbeh Mottaghi, and Ali Farhadi. Segan: Segmenting and generating the invisible. In *CVPR*, pages 6144–6153, 2018. 3
- [10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014. 1, 3, 5, 6
- [11] Paul Engstler, Andrea Vedaldi, Iro Laina, and Christian Rupprecht. Invisible stitch: Generating smooth 3d scenes with depth inpainting. *arXiv preprint arXiv:2404.19758*, 2024. 2, 3, 6, 7, 8
- [12] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018. 3
- [13] Yongtao Ge, Guangkai Xu, Zhiyue Zhao, Libo Sun, Zheng Huang, Yanlong Sun, Hao Chen, and Chunhua Shen. Geobench: Benchmarking and analyzing monocular geometry estimation models. *arXiv preprint arXiv:2406.12671*, 2024. 4
- [14] Ming Gui, Johannes S Fischer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. *arXiv preprint arXiv:2403.13788*, 2024. 2, 3, 4, 5, 7
- [15] Cheng-Yen Hsieh, Tarasha Khurana, Achal Dave, and Deva Ramanan. Tracking any object amodally. *arXiv preprint arXiv:2312.12433*, 2023. 3
- [16] Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T Freeman, David Salesin, Brian Curless, et al. Slide: Single image 3d photography with soft layering and depth-aware inpainting. In *ICCV*, pages 12518–12527, 2021. 3
- [17] Seong-Uk Jo, Du Yeol Lee, and Chae Eun Rhee. Occlusion-aware amodal depth estimation for enhancing 3d reconstruction from a single image. *IEEE Access*, 2024. 2, 3, 4, 6, 7
- [18] Abhishek Kar, Shubham Tulsiani, Joao Carreira, and Jitendra Malik. Amodal completion and size constancy in natural scenes. In *ICCV*, pages 127–135, 2015. 3
- [19] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, pages 9492–9502, 2024. 3
- [20] Lei Ke, Yu-Wing Tai, and Chi-Keung Tang. Deep occlusion-aware instance segmentation with overlapping bilayers. In *CVPR*, pages 4019–4028, 2021. 3
- [21] 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 *ICCV*, pages 4015–4026, 2023. 2, 4
- [22] Mykola Lavreniuk, Shariq Farooq Bhat, Matthias Muller, and Peter Wonka. Evp: Enhanced visual perception using inverse multi-attentive feature refinement and regularized image-text alignment. In *European Conference on Computer Vision Workshops (ECCVW)*, 2024. 3
- [23] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *arXiv preprint arXiv:2204.00987*, 2022. 1, 3, 4, 7
- [24] Zhenyu Li, Shariq Farooq Bhat, and Peter Wonka. Patch-fusion: An end-to-end tile-based framework for high-resolution monocular metric depth estimation. *arXiv preprint arXiv:2312.02284*, 2023. 3
- [25] Zhenyu Li, Zehui Chen, Xianming Liu, and Junjun Jiang. Depthformer: Exploiting long-range correlation and local information for accurate monocular depth estimation. *Machine Intelligence Research*, pages 1–18, 2023. 3, 4, 7
- [26] Ruoshi Liu and Carl Vondrick. Humans as light bulbs: 3d human reconstruction from thermal reflection. In *CVPR*, pages 12531–12542, 2023. 3
- [27] Alican Mertan, Damien Jade Duff, and Gozde Unal. Single image depth estimation: An overview. *Digital Signal Processing*, 123:103441, 2022. 1
- [28] Ege Ozguroglu, Ruoshi Liu, Dídac Surís, Dian Chen, Achal Dave, Pavel Tokmakov, and Carl Vondrick. pix2gestalt:

- Amodal segmentation by synthesizing wholes. In *CVPR*, pages 3931–3940. IEEE Computer Society, 2024. 2, 3, 4, 6, 7, 8
- [29] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, pages 10106–10116, 2024. 2, 4
- [30] 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. 3, 7, 8
- [31] Lu Qi, Li Jiang, Shu Liu, Xiaoyong Shen, and Jiaya Jia. Amodal instance segmentation with kins dataset. In *CVPR*, pages 3014–3023, 2019. 3
- [32] Siyuan Qiao, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. Vip-deeplab: Learning visual perception with depth-aware video panoptic segmentation. In *CVPR*, pages 3997–4008, 2021. 3
- [33] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, pages 12179–12188, 2021. 5
- [34] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3), 2022. 2, 3, 4, 5
- [35] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 4
- [36] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360monodepth: High-resolution 360deg monocular depth estimation. In *CVPR*, pages 3762–3772, 2022. 5
- [37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 3
- [38] Ahmed Rida Sekkat, Rohit Mohan, Oliver Sawade, Elmar Matthes, and Abhinav Valada. Amodalsynthdrive: A synthetic amodal perception dataset for autonomous driving. *IEEE Robotics and Automation Letters*, 2024. 2, 3, 4, 6
- [39] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, pages 8028–8038, 2020. 2, 3
- [40] Zeyi Sun, Ye Fang, Tong Wu, Pan Zhang, Yuhang Zang, Shu Kong, Yuanjun Xiong, Dahua Lin, and Jiaqi Wang. Alpha-clip: A clip model focusing on wherever you want. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13019–13029, 2024. 5
- [41] Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *CVPR*, pages 9763–9772, 2024. 3
- [42] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *NeurIPS*, 30, 2017. 3
- [43] Yufei Wang, Bo Li, Ge Zhang, Qi Liu, Tao Gao, and Yuchao Dai. Lrru: Long-short range recurrent updating networks for depth completion. In *CVPR*, pages 9422–9432, 2023. 3
- [44] Yufei Wang, Ge Zhang, Shaoqian Wang, Bo Li, Qi Liu, Le Hui, and Yuchao Dai. Improving depth completion via depth feature upsampling. In *CVPR*, pages 21104–21113, 2024.
- [45] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth completion. In *CVPR*, pages 4874–4884, 2024. 3
- [46] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. *arXiv preprint arXiv:2401.10891*, 2024. 1, 2, 3, 4
- [47] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 2, 3, 4, 7
- [48] Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. Amodal ground truth and completion in the wild. In *CVPR*, pages 28003–28013, 2024. 2, 3
- [49] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *CVPR*, pages 3784–3792, 2020. 2, 3
- [50] Xiaohang Zhan, Xingang Pan, Bo Dai, Ziwei Liu, Dahua Lin, and Chen Change Loy. Self-supervised scene de-occlusion. In *CVPR*, pages 3784–3792, 2020. 2
- [51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 7
- [52] Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *CVPR*, pages 18527–18536, 2023. 3
- [53] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5729–5739, 2023. 3
- [54] Yan Zhu, Yuandong Tian, Dimitris Metaxas, and Piotr Dollár. Semantic amodal segmentation. In *CVPR*, pages 1464–1472, 2017. 3