Exploiting Diffusion Prior for Generalizable Dense Prediction

Hsin-Ying Lee[1] Hung-Yu Tseng[2] Hsin-Ying Lee[3] Ming-Hsuan Yang[1,4]

[1]University of California, Merced [2]Meta [3]Snap Research [4]Yonsei University

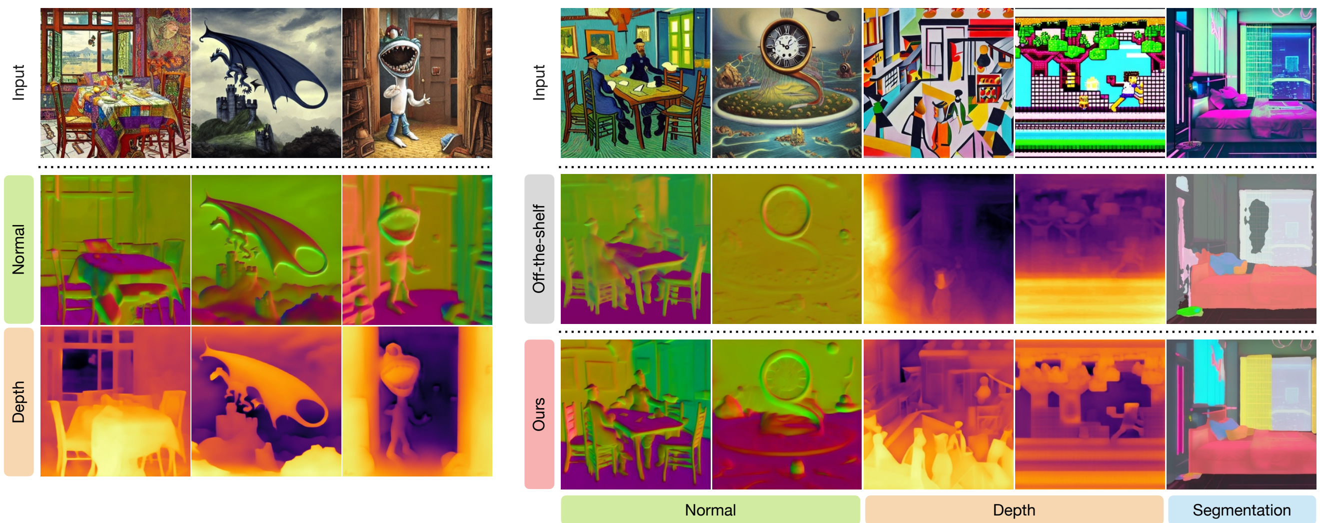


Figure 1. **Generalized dense prediction**. (*left*) We leverage the pre-trained text-to-image diffusion model [47] as a *prior* for various dense prediction tasks. (*right*) With only a small amount of labeled training data in a *limited* domain (i.e., 10K bedroom images with labels) for each task, our method performs favorably against SOTA predictors [5, 15, 26] on *arbitrary* images.

Abstract

Contents generated by recent advanced Text-to-Image (T2I) diffusion models are sometimes too imaginative for existing off-the-shelf dense predictors to estimate due to the immitigable domain gap. We introduce `DMP`*, a pipeline utilizing pre-trained T2I models as a prior for dense prediction tasks. To address the misalignment between deterministic prediction tasks and stochastic T2I models, we reformulate the diffusion process through a sequence of interpolations, establishing a deterministic mapping between input RGB images and output prediction distributions. To preserve generalizability, we use low-rank adaptation to fine-tune pre-trained models. Extensive experiments across five tasks, including 3D property estimation, semantic segmentation, and intrinsic image decomposition, showcase the efficacy of the proposed method. Despite limited-domain training data, the approach yields faithful estimations for arbitrary images, surpassing existing state-of-the-art algorithms. The code is available at* `https://github.com/shinying/dmp`.

1. Introduction

Text-to-image (T2I) diffusion models [11, 19, 47, 50] have achieved unprecedented progress on text-guided image generation, producing highly imaginative and realistic images from diverse and free-from textual descriptions. These advancements open up a new era of AI-aided content creation with applications spanning various domains [1, 22, 43, 45, 49, 53, 61]. However, the creativity of images generated by T2I models poses challenges for off-the-shelf dense (e.g., depth, normal, segmentation) prediction methods [5, 15, 26] due to the domain gap. For example, the ZoeDepth [5] approach fails to accurately predict the depth of the cubism painting shown in the sixth column in Figure 1. Such property predictions are vital for understanding high-level semantics of generated contents and can facilitate various downstream applications such as 3D imaging [54] and relighting [73].

Existing dense prediction models are typically trained on "real-world" images regardless of the training dataset scale. While these models aim for generalization, bridging the domain gap between real-world and T2I-generated images remains challenging, as we demonstrate in the right-hand side of Figure 1. A potential solution is to take advantage of the inherent generalizability of pre-trained T2I models, for example, by formulating dense prediction as image-to-image translation. However, while several recent efforts have been made to solve various image-to-image translations with pre-trained T2I models [7, 22, 43], we show in Section 4 that these methods are not directly applicable.

Leveraging pre-trained T2I models as a prior for dense prediction is challenging for two reasons. First, most dense prediction tasks are inherently *deterministic*, posing difficulties when adapting a pre-trained T2I model designed for stochastic text-to-image generation. Second, it is crucial to strike a balance between learning target tasks and retaining the inherent generalizability of pre-trained T2I models. In other words, learned dense predictors should generalize to arbitrary images from the training data in a limited domain.

In this paper, we propose `DMP` (**D**iffusion **M**odels as **P**riors) to leverage the pre-trained T2I model [47] as a prior for generalized dense prediction. To resolve the determinism-stochasticity misalignment, we introduce a deterministic mapping between the input RGB images and output prediction distributions. Specifically, we reformulate the diffusion process as a chain of interpolations between input RGB images and their corresponding output signals, where the importance of input images gradually increases over the diffusion process. The reverse diffusion (i.e., known as denoising or generation in original T2I) process becomes a series of transformations that progressively synthesize desired output signals from input images. Without randomization, such as Gaussian noise imposed, the mapping is entirely deterministic. In addition, to retain the generalizability of the pre-trained T2I model while learning target tasks, we use low-rank adaptation [24] to fine-tune the pre-trained model with the aforementioned deterministic diffusion process for each dense prediction task. Figure 1 demonstrates the generalization ability of the proposed method on the deterministic normal, depth, and segmentation prediction problems.

We conduct extensive quantitative and qualitative experiments on five dense prediction tasks to evaluate the proposed `DMP` approach: 3D property estimation (depth, normal), semantic segmentation, and intrinsic image decomposition (albedo, shading). We show that with only a small amount of limited-domain training data (i.e., 10K bedroom training images with labels), the proposed method can provide faithful estimations of the in-domain and unseen images, especially those that the existing SOTA algorithms struggle to handle effectively. We summarize the contributions as follows:

- We propose `DMP`, an approach leveraging the pre-trained T2I model as a prior for dense prediction tasks.
- We design an image-to-prediction diffusion process that adapts the stochastic T2I model for deterministic dense prediction problems.
- We use five dense prediction tasks to validate that the proposed method obtains faithful estimation on arbitrary images despite training with a small amount of data in a limited domain.

2. Related Work

Diffusion models. Diffusion models estimate a target data distribution by modeling the transition from a noisy version of the distribution [23, 57]. Recently, diffusion models have shown unprecedented quality in the text-to-image [47, 50] setting by training on large-scale datasets [52]. The advancements unleash various text-guided image manipulation applications [1, 7, 9, 22, 40, 43, 45, 53]. The stochasticity in the generation process, a preferred property in most applications, derives from the initial noise and additional noise added during the denoising process. However, the dense prediction problems are usually deterministic. Some recent methods adopt deterministic sampling algorithms [30, 35, 58] and reformulate the diffusion and generative process by α-blending and de-blending [21]. Though the mapping between initial noise and outputs in target distributions is deterministic, the correlation between each noise and output sample is stochastic. If adapted to deterministic tasks, the model may generate high-quality outputs unfaithful to input images.

Image-to-image translation. The task aims to learn a mapping between two visual domains. Early efforts [8, 25, 31, 42, 79, 80] mostly make use of generative adversarial networks and cycle consistency loss to learn the translation from scratch. With StyleGAN [28, 29] showing high-quality synthesis on certain categories, some methods [46, 63] seek to achieve image-to-image translation using pre-trained StyleGAN by training an additional encoder. However, the translation is limited to certain categories. Recently, following the success of large-scale text-to-image models, attempts have been made to perform image-to-image translation with pre-trained diffusion models [7, 22, 43, 65]. These methods, however, are not directly applicable to the dense prediction task.

Fine-tuning text-to-image diffusion models. In addition to image-to-image translation, pre-trained diffusion models are adopted to take additional modalities [33, 75], be customized on certain objects [16, 48] or styles [76], and synthesize videos [17, 20, 36]. These methods train additional zero-initialized layers [75], manipulate attention modules [17, 33], learn a token embedding [16, 76], or learn parameter offsets with low-rank matrices [24]. In this work, we adopt the low-rank adaptation [24] to fine-tune only parameter offsets of attention layers.

Generative prior for dense prediction. Prior work has leveraged pre-trained generative models as a prior for other tasks, such as representation learning [13], as the latent features of pre-trained generative models are found to be rich in semantics [64, 70, 72]. Bhattad et al. [6] manipulates style latents of StyleGAN [27] and reveals its learned ability to estimate image properties, but the generalizability is limited. Some works reuse [2, 18, 71, 77] or merge [62] latent features of pre-trained denoising U-Nets to perform segmentation and depth estimation. Others [67, 68] transform generation models into multi-task generalists by standardizing the outputs of tasks as images. In this work, instead of developing a specific approach for a particular task, we focus on analyzing the potential of pre-trained models as a prior for general dense prediction through a universal transferring framework.

3. Method

Our goal is to leverage the pre-trained T2I diffusion model as a prior to learn a dense prediction task from a set of labeled training data $\mathcal{D} = \{(x^i, y^i)\}_{i=1}^{n}$, where x^i denotes the input image, y^i indicates the corresponding output (e.g., depth map), and $x^i, y^i \in \mathbb{R}^{H\times W\times 3}$. We first describe the text-to-diffusion models used as the prior in Section 3.1. Then we introduce the proposed `DMP` approach in Section 3.2.

3.1. Text-to-Image Diffusion Models

We use the pre-trained T2I latent diffusion model [47] as the prior. It consists of an autoencoder and a U-Net. The autoencoder converts between an image $y \in \mathbb{R}^{H\times W\times 3}$ and its latent feature $\tilde{y} \in \mathbb{R}^{h\times\, w\times c}$, where $(h, w) = (H/8, W/8)$ and c represents the channel size of latent features. Since we do not modify the autoencoder in the proposed approach, we use y to represent the latent feature $\tilde{y}$ to simplify the annotation in this paper.

The U-Net model takes as input a text description and learns to reverse the following diffusion process that gradually turns the image y into noise map y_T:

$$y_t = \sqrt{\bar{\alpha}_t}y + \sqrt{1-\bar{\alpha}_t}\epsilon_t \qquad \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{1}$$

where $t = [1, \cdots, T]$, and $\bar{\alpha}_t$ is the noise schedule [23].

3.2. Leveraging Diffusion Prior

There are two challenges to leverage (*i.e*., fine-tune) the pre-trained T2I approach for estimating the pixel-level output (*e.g*., depth) y from an input image x: 1) determinism-stochasticity misalignment and 2) generalizability. We introduce the solutions to tackle these two issues as follows.

Deterministic diffusion. The diffusion process described in Eq. (1) is designed specifically for stochastic image generation. However, the mapping between input images and outputs in dense prediction problems is typically deterministic. We observe that directly applying the diffusion process in Eq. (1) to the dense prediction tasks introduces unnecessary variation in outputs that leads to apparent artifacts. Therefore, we use the blending strategy [21, 34] and re-design the diffusion process as follows. Instead of converting from noise maps to images in the conventional T2I method, the diffusion process in our `DMP` approach directly maps between the input image x and output y. As illustrated in Figure 2, the proposed diffusion process is formulated as

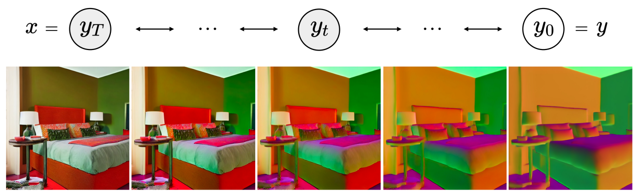


Figure 2. **Deterministic diffusion process.** We formulate the diffusion process as a chain of interpolations between an input image x and output y. The U-Net model is fine-tuned to gradually transform the input x to the desired dense prediction y.

$$y_t = \sqrt{\bar{\alpha}_t}y + \sqrt{1-\bar{\alpha}_t}x \qquad t = [1, \cdots, T]. \tag{2}$$

As we can see from Figure 2 and Eq. (2), the proposed scheme gradually increases the weight of the input image x over the diffusion process. This can be considered as progressively morphing the output y into the input image x via interpolation. As a result, we can fine-tune the U-Net model to reverse the diffusion process that interactively "demorphs" the input image x and gives the final prediction y.

Parameterization. We explore various parameterizations for the U-Net model to make different predictions. As discussed in Section 4.5, we empirically find that the v-prediction [51] works well for the dense prediction tasks. Specifically, the U-Net model v_θ is fine-tuned using the following mean square loss:

$$L_{\text{DMP}} = \mathbb{E}_{(x,y),t}\big[\|(\sqrt{\bar{\alpha}_t}x - \sqrt{1-\bar{\alpha}_t}y) - v_\theta(y_t, t)\|_2^2\big], \tag{3}$$

where $v_\theta(y_t, t)$ is the U-Net prediction. The reverse diffusion process can then be formulated as

$$\begin{aligned} y_{t-1} = &\sqrt{\bar{\alpha}_{t-1}}(\sqrt{\bar{\alpha}_t}y_t - \sqrt{1-\bar{\alpha}_t}v_\theta(y_t, t)) \\ &+ \sqrt{1-\bar{\alpha}_{t-1}}x \qquad t = [T, \cdots, 1], \end{aligned} \tag{4}$$

where $y_T = x$ and y_0 is the desired output.

U-Net fine-tuning. To learn the target tasks while retaining the inherent generalization ability of the pre-trained T2I model, we use the low-rank approximation [24] scheme to fine-tune all the attention layers in the U-Net model to minimize the objective described in Eq. (3).

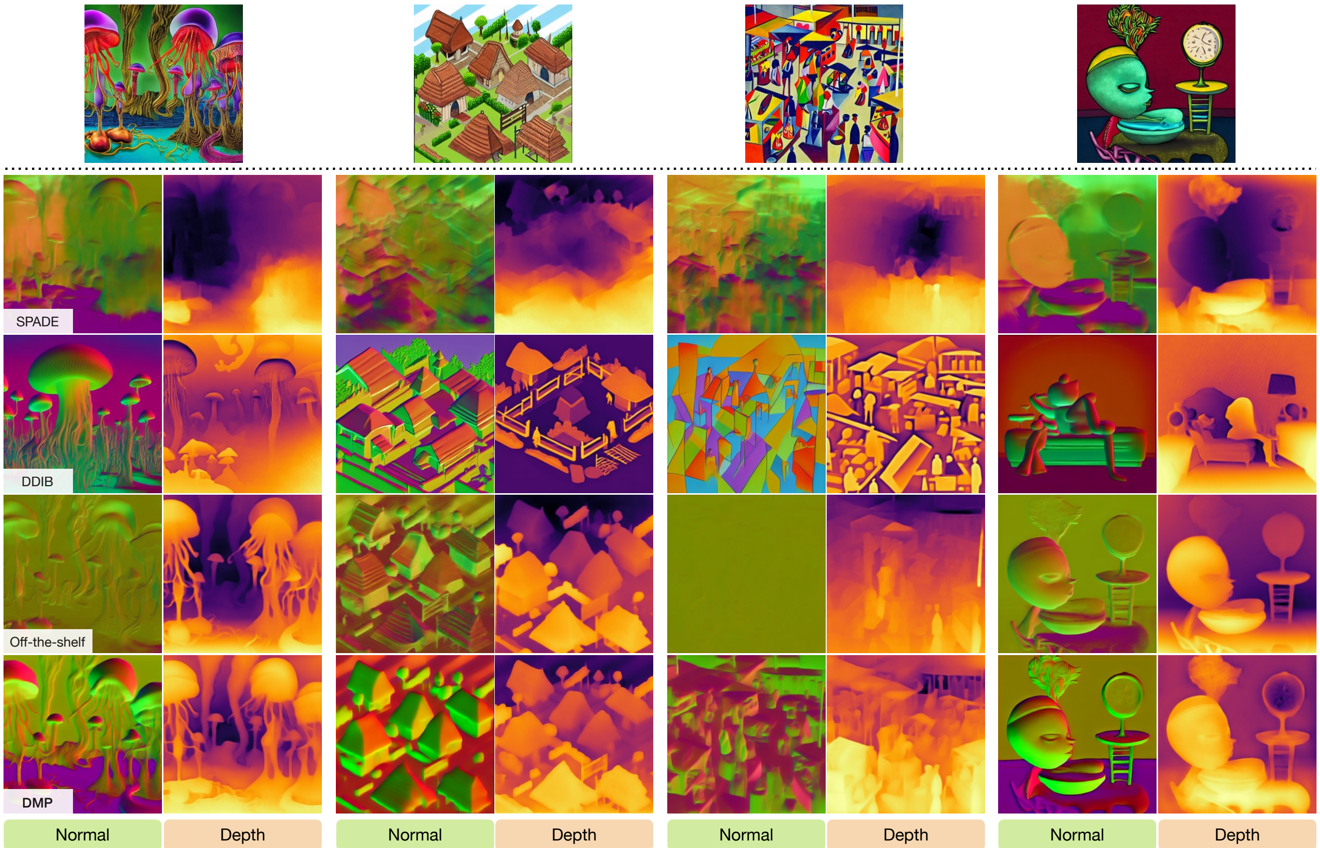


Figure 3. **3D property estimation of *arbitrary* input images.** The first row shows the input images, while the remaining rows present the normals and depth estimated by different approaches. The proposed `DMP` method gives faithful estimation, even on the images where the off-the-shelf [5, 26] schemes fail to handle.

4. Experimental Results

4.1. Experiment Setup

We evaluate the proposed `DMP` approach using five dense prediction tasks, including 3D property estimation (i.e., depth and normals), semantic segmentation, and intrinsic image decomposition (i.e., albedo and shading).

Datasets. To better analyze generalizability by varying training and test data domains, we conduct the evaluation with synthetic images. We first generate diverse text descriptions using a large language model [39] by filling a *keyword* in a prompt template modified from the one used in Parmar et al. [44]. We then use a text-to-image diffusion model [47] to synthesize the images according to the text descriptions. Second, we use the following off-the-shelf predictors to generate the *pseudo* ground truth for each image: Omnidata v2 [26] for surface normals, ZoeDepth [5] for monocular depth, EVA-2 [15] for semantic segmentation, and PIE-Net [12] for intrinsic image decomposition (albedo, shading). Finally, we follow the protocol used in Bhattad et al. [6] to generate a set of training data, and three sets of test data:

- Training set: We generate 10K labeled images using the keyword "bedroom".
- In-domain test set: We generate 2K labeled images using the keyword "bedroom".
- Out-of-domain test set: We use the 409 category names of the indoor scenes in the SUN dataset [69] as the keywords to generate 2K labeled images. The set is considered to be out-of-domain compared to the training set. Nevertheless, the off-the-shelf models that provide the pseudo ground truth still work well since the images belong to normal indoor scenes.
- *Arbitrary* test set: We use random textual descriptions to generate the images. Since the generated images are almost free-form, the off-the-shelf models cannot provide proper predictions. Therefore, we consider the off-the-shelf approaches as compared methods and present only the visual comparisons.

Note that we present the quantitative results only on the in-domain test set for semantic segmentation due to the semantic label set constraint. In addition, we show qualitative

Table 1. **Quantitative comparisons on 3D property estimation.** We compute the metrics using the estimated results and the pseudo ground truth generated by the off-the-shelf predictors.

	Normal				Depth					
	In-domain		Out-of-domain		In-domain			Out-of-domain		
	L1↓	Ang↓	L1↓	Ang↓	REL↓	$\delta\uparrow$	RMSE↓	REL↓	$\delta\uparrow$	RMSE↓
SPADE [41]	<u>0.0708</u>	<u>0.1635</u>	<u>0.1268</u>	<u>0.2833</u>	<u>0.2132</u>	0.4961	<u>0.1379</u>	<u>0.3587</u>	<u>0.3190</u>	<u>0.2554</u>
DRIT++ [32]	0.0784	0.1723	0.1350	0.3006	0.3792	0.2458	0.2134	0.4374	0.2585	0.3212
SDEdit [38]	0.2599	0.5087	0.2675	0.5293	0.4656	0.3533	0.3240	0.6640	0.2495	0.3382
DDIB [60]	0.1849	0.4210	0.2271	0.4847	0.3087	<u>0.5130</u>	0.2367	0.6275	0.2788	0.3120
IP2P (hard) [7]	0.3017	0.5468	0.3168	0.5757	0.4834	0.3235	0.3358	0.6450	0.2252	0.3461
IP2P (learned) [7]	0.3550	0.7181	0.3397	0.6836	0.3965	0.3302	0.3494	0.5182	0.2664	0.3261
VISII [40]	0.2081	0.4386	0.2448	0.4895	0.3498	0.4405	0.2912	0.5364	0.2855	0.3181
DMP	**0.0514**	**0.1156**	**0.0872**	**0.1886**	**0.1072**	**0.8861**	**0.1020**	**0.2117**	**0.6395**	**0.1360**

comparisons of bedroom images using diverse styles (e.g., cyberpunk, comic) to understand the generalization ability.

Compared methods. We compare our method with the GAN-based image translation methods SPADE [41] and DRIT++ [32, 37]. These models are trained from scratch using the training set (i.e., 10K labeled bedroom images). We also include for comparison the following approaches that leverage the pre-trained T2I model as the prior:

- **SDEdit** [38]: We fine-tune using the training label images $\{y^i\}_{i=1}^{10\text{K}}$ with the standard diffusion process in Eq. (1). Then we follow the original SDEdit approach that adds the noise to an input x and uses the fine-tuned model to de-noise for generating the output y.
- **DDIB** [60]: We use the same fine-tuned model in SDEdit and adopt the DDIB method to predict the output y from the input x.
- InstructPix2Pix (IP2P) [7]: We evaluate two versions. In **IP2P (hard)**, we use the pre-defined instructions such as “make it into the corresponding depth map” as the input to the model for inference. For the second version **IP2P (learned)**, we optimize the token $*$ in the input text “make it into $*$” using the training set.
- **VISII** [40]: We use their approach for fine-tuning with the training set and inference.

4.2. 3D Property Estimation

Surface normals [10, 14] and depth [56, 59] are crucial to 3D visual applications such as 3D reconstruction [74] and autonomous driving [66]. To evaluate normal prediction, we use the average L1 distance and average angular error Ang. For monocular depth, given the ground truth depth y^i and predicted depth $\hat{y}^i$, we use the average relative error $\mathrm{REL} = \frac{1}{M}\sum_{i=1}^{M}|y^i - \hat{y}^i|/y^i$, percentage of pixels δ where $\max(y^i/\hat{y}^i, \hat{y}^i/y^i) < 1.25$, and the root mean square error RMSE of the relative depth. Specifically, we normalize the ground-truth and predicted depth maps separately to be in the range of $[0, 1]$ as the relative depth.

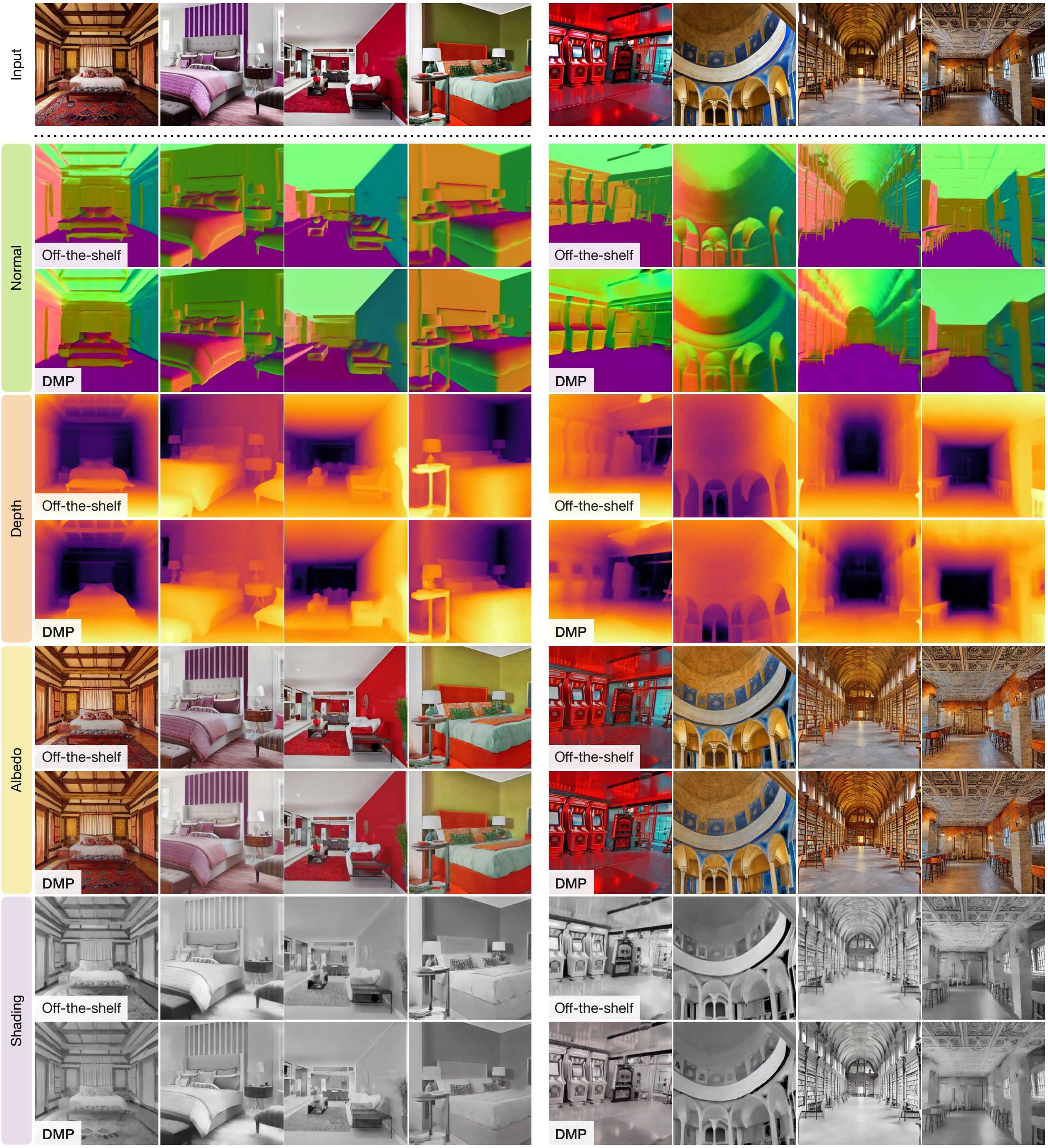
The quantitative comparisons are presented in Table 1, and qualitative results are shown in Figure 4. The proposed approach performs favorably against the compared algorithms in terms of accuracy and generalization ability. Although the REL and δ scores reported for our method degrade from the in-domain to out-of-domain test sets, the RMSE score of the relative depth remains the same. This is due to the scene scale difference between the training images (bedroom) and test images (larger spaces such as sports fields). The RMSE measurement is computed based on relative depth, which is more resistant to the scene scale change. Consistent with the quantitative results, we observe that the depth predicted by our method is in the correct order for the out-of-domain images in Figure 4.

Generalization ability. We demonstrate the generalizability of the proposed approach in Figure 3, where we use *arbitrary* images as the inputs. Although our method is fine-tuned with only 10K bedroom images with labels, it faithfully estimates the 3D property even on those images that the off-the-shelf methods fail to handle.

4.3. Semantic Segmentation

Semantic segmentation [78] is a fundamental visual understanding task. Since the prediction is categorical, we use a simple conversion for regression models to predict discrete labels. We first generate ground-truth labels using the off-the-shelf EVA-2 [15] model. The label maps are then transformed into color maps where each category has different colors. The training and inference of the diffusion model are conducted using the color maps (in the RGB space). During inference, the predicted color maps are converted to categorical label maps by assigning each pixel to its nearest category in the color space.

We report the intersection over union (IoU) and accuracy to measure the in-domain performance in Table 2. The prediction by the proposed scheme is reasonable across all categories compared to the existing methods. Figure 5 demonstrates the results of our method and pseudo ground truth with the in- and out-of-domain (i.e., bedroom images in di-



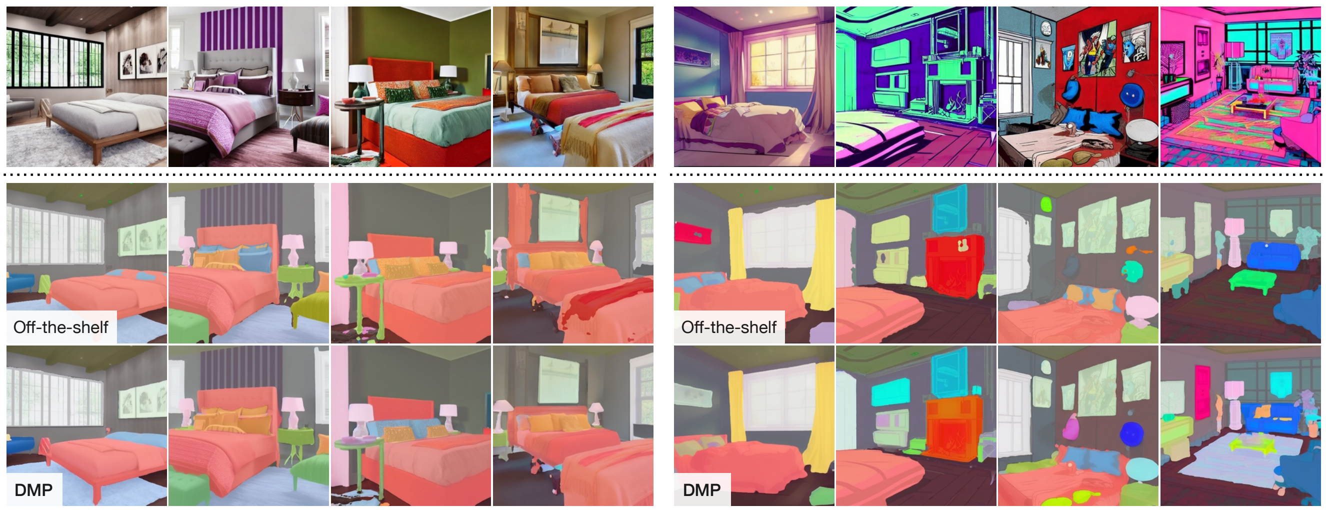
(a) In-domain (b) Out-of-domain

Figure 4. **Qualitative results.** The first row shows the input images. In the following, every two rows show the results predicted by the off-the-shelf predictors (which we considered as pseudo ground truth) and those by the proposed method.

verse styles) input images. Particularly in out-of-domain examples, our model gives favorable predictions compared to the off-the-shelf approach, e.g., the painting of the first, the curtain of the second, the window of the third, and the carpet of the fourth images in Figure 5 (b). This validates our idea that leverages the pre-trained T2I diffusion model as the prior for better generalization.

Table 2. **Quantitative comparisons on semantic segmentation.** We compute the metrics using the estimated results and the pseudo ground truth generated by the off-the-shelf predictors.

	Bed		Pillow		Lamp		Window		Painting		Mean	
	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑
SPADE [41]	0.8677	0.6370	0.5861	0.3473	0.3659	0.2084	0.6925	0.5627	0.5249	0.3826	0.6074	0.4276
DRIT++ [32]	0.8485	0.4587	0.2427	0.1435	0.1218	0.0776	0.3023	0.2414	0.2579	0.2114	0.3546	0.2265
SDEdit [38]	0.0958	0.0901	0.3824	0.0864	0.1522	0.0651	0.4501	0.2593	0.1333	0.0746	0.2428	0.1151
DDIB [60]	0.3984	0.3040	0.2256	0.0637	0.1630	0.0593	0.4741	0.2896	0.1728	0.0881	0.2868	0.1609
IP2P (learned) [7]	0.0714	0.0620	0.0086	0.0042	0.0228	0.0116	0.3532	0.1699	0.0386	0.0192	0.0989	0.0534
VISII [40]	0.0060	0.0059	0.0261	0.0136	0.0014	0.0011	0.2576	0.1772	0.0013	0.0012	0.0585	0.0398
DMP	**0.8947**	**0.8506**	**0.5871**	**0.3645**	**0.6399**	**0.4414**	**0.8338**	**0.7335**	**0.7490**	**0.6735**	**0.7409**	**0.6127**



(a) In-domain (b) Out-of-domain

Figure 5. **Qualitative results on semantic segmentation.** The first, second, and third rows respectively show the input images, pseudo ground truth predicted by an off-the-shelf model, and our results. The out-of-domain samples in (b) are bedroom images in diverse styles.

Table 3. **Quantitative comparisons on intrinsic image decomposition.** We compute the metrics to measure the difference between the estimated results and the pseudo ground truth created by the off-the-shelf predictors.

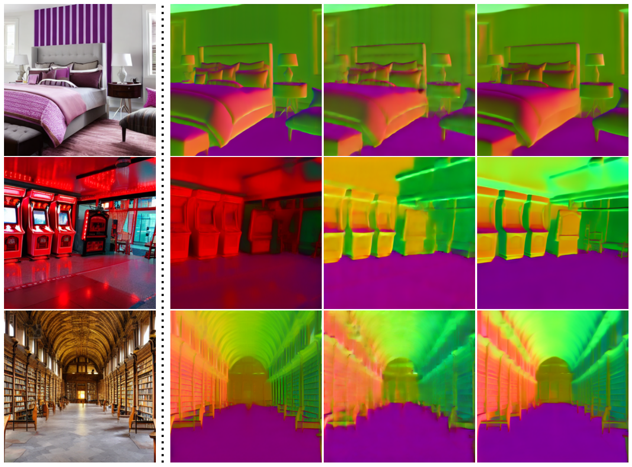
	Albedo		Shading	
	In	Out	In	Out
SPADE [41]	**0.0021**	**0.0030**	**0.0031**	**0.0040**
DRIT++ [32]	0.0296	0.0392	0.0309	0.0408
SDEdit [38]	0.0375	0.0471	0.0501	0.0671
DDIB [60]	0.0411	0.0443	0.0403	0.0557
IP2P (hard) [7]	0.0329	0.0479	0.0361	0.0421
IP2P (learned) [7]	0.0215	0.0250	0.0290	0.0309
VISII [40]	0.0145	0.0246	0.0275	0.0285
DMP	0.0041	0.0064	0.0051	0.0070

Table 4. **Effect of different parameterizations.** We fine-tune the U-Net model to predict different signals in each reverse diffusion step to get the final outputs: predicting the input image x, predicting the output y, and v-prediction described in Eq. (3). The experiment is conducted on the surface normal prediction task.

	In-domain		Out-of-domain	
	L1↓	Ang↓	L1↓	Ang↓
Predicting x	0.0736	0.1629	0.1319	0.2764
Predicting y	0.0590	0.1291	0.0888	0.1914
v-prediction	**0.0514**	**0.1156**	**0.0872**	**0.1886**

4.4. Intrinsic Image Decomposition

Intrinsic image decomposition [4] recovers albedo and shading properties from RGB images. It facilitates applications such as recoloring [3] and relighting [55]. Similar to PIE-Net [12], we use the mean square error MSE as the evaluation metric to compute the distance between the predicted and pseudo ground-truth estimations. We showcase the qualitative results in Figure 4 and quantitative comparisons in Table 3. The reported errors of most methods are an order of magnitude larger than ours. Notably, while our method and SPADE work well in both in- and out-of-



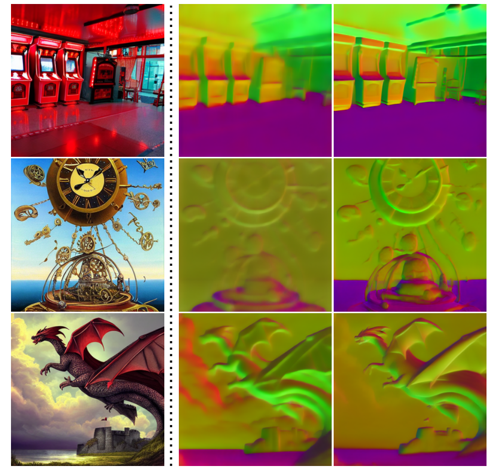


Figure 6. **Qualitative comparisons between various parameterizations.** We fine-tune the U-Net model to predict different signals in each reverse diffusion step to get the final output: predicting the input image x, predicting the output y, and v-prediction described in Eq. (3).

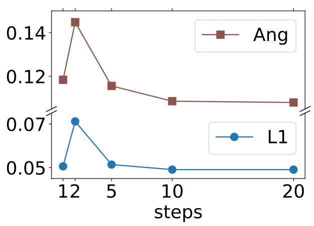


Figure 7. **Effect of varying the numbers of diffusion steps.** We report the in-domain performance on the surface normal prediction task. In the case of single-step, we train the U-Net model to directly predict outputs from input images.

Figure 8. **Single vs. multiple diffusion steps.** In the single-step approach, we train the U-Net network to directly predict outputs from input images. The single-step approach does not generalize to unseen images well.

domain settings, we found our method less influenced by artifacts in pseudo ground truth and to offer more reasonable estimations. More details are provided in the supplementary document.

4.5. Ablation Study

Parameterization. We study the effect of employing various parameterizations. Specifically, we fine-tune the U-Net model to make different predictions in each reverse diffusion step for obtaining the final prediction: 1) predicting the input image x, predicting the output y (similar to x_0-prediction in standard diffusion models), and v-prediction described in Eq. (3). We formulate each parameterization in detail in the supplementary document. The quantitative results are shown in Table 4 and the qualitative comparisons are presented in Figure 6. Predicting the input image x generates accurate results in the in-domain setting. However, it fails to generalize to unseen domains. On the other hand, predicting the output y demonstrates preferred generalization capability, but tends to produce blurry results with few details. We choose to use the v-prediction approach as it produces accurate results of arbitrary images.

Number of diffusion steps. We analyze the performance of our method with different numbers of diffusion steps in the inference stage. The quantitative comparisons on the normal prediction are presented in Figure 7. Using 5 steps strikes a good balance between inference speed and accuracy. Furthermore, we study an extreme case of using a single step. We train the U-net to directly predict the output y from the input image x. Although it shows competitive performance in the in-domain setting according to Figure 7, we find that the performance degrades significantly with the *arbitrary* input images, as examples shown in Figure 8. Considering the generalizability, inference speed, and accuracy, we use 5 generation steps for all the other experiments.

5. Conclusion

In this work, we leverage a pre-trained T2I diffusion model for generalizable dense prediction. The core of our method is the design of the deterministic diffusion process that adapts the stochastic T2I framework for the deterministic prediction tasks. With low-rank approximation, the proposed approach learns the target tasks while retaining the inherent generalization ability of the T2I model. We show that with only a small number of labeled training data in a limited domain (i.e., 10K bedroom images), our `DMP` scheme makes faithful predictions on *arbitrary* images. We believe that this work establishes a foundation for achieving ultimately-generalizable visual understanding in the future.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 1, 2

[2] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *ICLR*, 2022. 3

[3] Shida Beigpour and Joost van de Weijer. Object recoloring based on intrinsic image estimation. In *ICCV*, 2011. 7

[4] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM TOG*, 33(4), 2014. 7

[5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arxiv:2302.12288*, 2023. 1, 4

[6] Anand Bhattad, Daniel McKee, Derek Hoiem, and D. A. Forsyth. Stylegan knows normal, depth, albedo, and more. *arXiv preprint arXiv:2306.00987*, 2023. 3, 4

[7] Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, 2023. 2, 5, 7

[8] Yu-Jie Chen, Shin-I Cheng, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Vector quantized image-to-image translation. In *ECCV*, 2022. 2

[9] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Adaptively-realistic image generation from stroke and sketch with diffusion model. In *WACV*, 2023. 2

[10] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 5

[11] Xiaoliang Dai, Ji Hou, Chih-Yao Ma, Sam Tsai, Jialiang Wang, Rui Wang, Peizhao Zhang, Simon Vandenhende, Xiaofang Wang, Abhimanyu Dubey, Matthew Yu, Abhishek Kadian, Filip Radenovic, Dhruv Mahajan, Kunpeng Li, Yue Zhao, Vladan Petrovic, Mitesh Kumar Singh, Simran Motwani, Yi Wen, Yiwen Song, Roshan Sumbaly, Vignesh Ramanathan, Zijian He, Peter Vajda, and Devi Parikh. Emu: Enhancing image generation models using photogenic needles in a haystack. *arXiv preprint arXiv:2309.15807*, 2023. 1

[12] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *CVPR*, 2022. 4, 7

[13] Jeff Donahue and Karen Simonyan. Large scale adversarial representation learning. In *NeurIPS*, 2019. 3

[14] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, 2021. 5

[15] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023. 1, 4, 5

[16] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023. 2

[17] Michal Geyer, Omer Bar-Tal, Shai Bagon, and Tali Dekel. Tokenflow: Consistent diffusion features for consistent video editing. *arXiv preprint arXiv:2307.10373*, 2023. 2

[18] Rui Gong, Martin Danelljan, Han Sun, Julio Delgado Mangas, and Luc Van Gool. Prompting diffusion representations for cross-domain semantic segmentation. *arXiv preprint arXiv:2307.02138*, 2023. 3

[19] Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Josh Susskind, and Navdeep Jaitly. Matryoshka diffusion models. *arXiv preprint arXiv:2310.15111*, 2023. 1

[20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023. 2

[21] Eric Heitz, Laurent Belcour, and Thomas Chambon. Iterative α-(de)blending: A minimalist deterministic diffusion model. In *ACM SIGGRAPH*, 2023. 2, 3

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. Prompt-to-prompt image editing with cross-attention control. In *ICLR*, 2023. 1, 2

[23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[24] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 3

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

[26] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *CVPR*, 2022. 1, 4

[27] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 3

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *CVPR*, 2020. 2

[30] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022. 2

[31] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *ECCV*, 2018. 2

[32] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Kumar Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation viadisentangled representations. *IJCV*, pages 1–16, 2020. 5, 7

[33] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In *CVPR*, 2023. 2

[34] Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos A Theodorou, Weili Nie, and Anima Anandkumar. I^2sb: Image-to-image schrödinger bridge. In *ICML*, 2023. 3

[35] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan LI, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. In *NeurIPS*, 2022. 2

[36] Aniruddha Mahapatra, Aliaksandr Siarohin, Hsin-Ying Lee, Sergey Tulyakov, and Jun-Yan Zhu. Text-guided synthesis of eulerian cinemagraphs. *SIGGRAPH Asia*, 2023. 2

[37] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, 2019. 5

[38] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 5, 7

[39] Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *ACL*, pages 15991–16111, 2023. 4

[40] Thao Nguyen, Yuheng Li, Utkarsh Ojha, and Yong Jae Lee. Visual instruction inversion: Image editing via visual prompting. In *NeurIPS*, 2023. 2, 5, 7

[41] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 5, 7

[42] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 2

[43] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH*, 2023. 1, 2

[44] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. In *ACM SIGGRAPH*, 2023. 4

[45] Tanzila Rahman, Hsin-Ying Lee, Jian Ren, Sergey Tulyakov, Shweta Mahajan, and Leonid Sigal. Make-a-story: Visual memory conditioned consistent story generation. In *CVPR*, 2023. 1, 2

[46] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in style: a stylegan encoder for image-to-image translation. In *CVPR*, 2021. 2

[47] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4

[48] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 2

[49] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, 2023. 1

[50] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022. 1, 2

[51] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 3

[52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *NeurIPS Datasets and Benchmarks Track*, 2022. 2

[53] Yujun Shi, Chuhui Xue, Jiachun Pan, Wenqing Zhang, Vincent YF Tan, and Song Bai. Dragdiffusion: Harnessing diffusion models for interactive point-based image editing. *arXiv preprint arXiv:2306.14435*, 2023. 1, 2

[54] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020. 1

[55] Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. Neural face editing with intrinsic image disentangling. In *CVPR*, 2017. 7

[56] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5

[57] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 2

[58] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *ICLR*, 2021. 2

[59] Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 5

[60] Xuan Su, Jiaming Song, Chenlin Meng, and Stefano Ermon. Dual diffusion implicit bridges for image-to-image translation. In *ICLR*, 2023. 5, 7

[61] Yu Takagi and Shinji Nishimoto. High-resolution image reconstruction with latent diffusion models from human brain activity. In *CVPR*, 2023. 1

[62] Junjiao Tian, Lavisha Aggarwal, Andrea Colaco, Zsolt Kira, and Mar Gonzalez-Franco. Diffuse, attend, and segment: Unsupervised zero-shot segmentation using stable diffusion. *arXiv preprint arXiv:2308.12469*, 2023. 3

[63] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM TOG*, 2021. 2

[64] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing gans for one-shot semantic part segmentation. In *CVPR*, 2021. 3

[65] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 2

[66] Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *3DV*, 2017. 5

[67] Xinlong Wang, Wen Wang, Yue Cao, Chunhua Shen, and Tiejun Huang. Images speak in images: A generalist painter for in-context visual learning. In *CVPR*, 2023. 3

[68] Zhendong Wang, Yifan Jiang, Yadong Lu, Yelong Shen, Pengcheng He, Weizhu Chen, Zhangyang Wang, and Mingyuan Zhou. In-context learning unlocked for diffusion models. In *NeurIPS*, 2023. 3

[69] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *IJCV*, 119:3–22, 2016. 4

[70] Jianjin Xu and Changxi Zheng. Linear semantics in generative adversarial networks. In *CVPR*, 2021. 3

[71] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *CVPR*, 2023. 3

[72] Yinghao Xu, Yujun Shen, Jiapeng Zhu, Ceyuan Yang, and Bolei Zhou. Generative hierarchical features from synthesizing images. In *CVPR*, 2021. 3

[73] Yu-Ying Yeh, Zhengqin Li, Yannick Hold-Geoffroy, Rui Zhu, Zexiang Xu, Miloš Hašan, Kalyan Sunkavalli, and Manmohan Chandraker. Photoscene: Photorealistic material and lighting transfer for indoor scenes. In *CVPR*, 2022. 1

[74] Amir R. Zamir, Alexander Sax, William Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *CVPR*, 2018. 5

[75] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 2

[76] Yuxin Zhang, Nisha Huang, Fan Tang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. Inversion-based style transfer with diffusion models. In *CVPR*, 2023. 2

[77] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *ICCV*, 2023. 3

[78] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 5

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

[80] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *NeurIPS*, 2017. 2