Exploiting Diffusion Prior for Generalizable Dense Prediction

Supplementary Material

A. Parametrizations

As described in Section 3.2, we empirically find that parametrizing the U-Net model through estimating v-prediction [16] performs favorably against predicting inputs or outputs. We detail the formulation of predicting inputs and outputs as follows. The U-Net model $\hat{x}_\theta$ predicting inputs is fine-tuned with the mean square loss:

$$L = \mathbb{E}_{(x,y),t}\left[\|x - \hat{x}_\theta(y_t, t)\|_2^2\right], \tag{1}$$

and the reverse diffusion process is formulated as

$$\begin{aligned} y_{t-1} &= \sqrt{\bar{\alpha}_{t-1}}\left(\frac{y_t - \sqrt{1-\bar{\alpha}_t}\hat{x}_\theta(y_t,t)}{\sqrt{\bar{\alpha}_t}}\right) \\ &+ \sqrt{1-\bar{\alpha}_{t-1}}\hat{x}_\theta(y_t,t) \qquad t = [T, \cdots, 1], \end{aligned} \tag{2}$$

The U-Net model $\hat{y}_\theta$ predicting outputs is optimized with the loss function:

$$L = \mathbb{E}_{(x,y),t}\left[\|y - \hat{y}_\theta(y_t, t)\|_2^2\right], \tag{3}$$

and the reverse diffusion process is

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

B. Additional Experimental Results

B.1. Reliability of Off-the-Shelf Estimators

We indicate that the off-the-shelf estimators are not always reliable, especially the approach for intrinsic image decomposition. We demonstrate with the example of albedo estimation in Figure 1 that the off-the-shelf approach generates apparent artifacts in shadow areas, such as corners or floors under beds. The approach fails to recover the correct albedo but instead generates black patches. Consequently, SPADE also learns this pattern, but our model tends to correct artifacts by performing accurate estimation, manifesting the ability of generalization.

B.2. Real-World Evaluation

NYU Depth v2. Following Ke et al. [6], we evaluate our method on NYU Depth v2 [18] according to the protocol of affine-invariant depth evaluation [13]. We generate prompts with BLIP-2 [7] to use the model trained on synthetic bedroom images. We scale and shift the depth predictions to align ground truths by solving least-square fitting. The comparison against other approaches is shown in Table 1. DMP performs comparably with some previous methods trained with large-scale data.

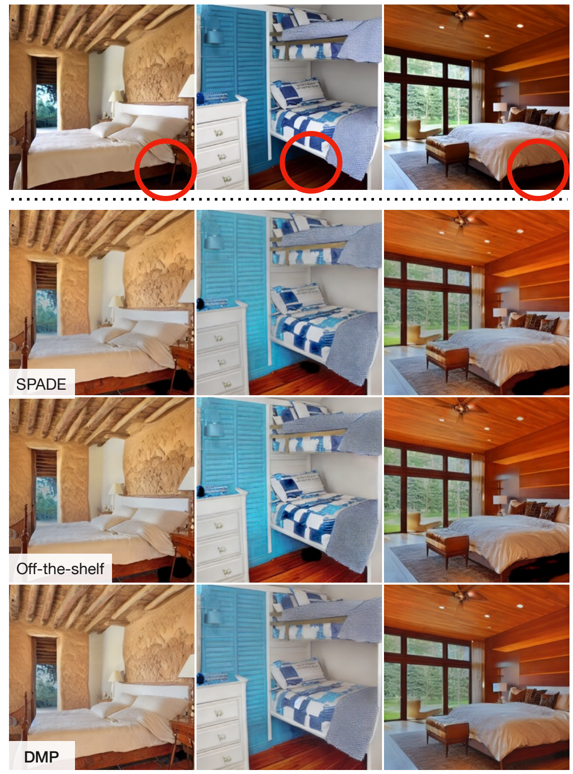


Figure 1. **Qualitative comparisons on albedo estimation**. SPADE [9] and the off-the-shelf approaches generate artifacts in dark areas.

Table 1. Comparison of performance on NYU Depth v2 [18].

	# Training Samples		NYU v2	
	Real	Synthetic	REL↓	$\delta\uparrow$
MiDaS [13]	2M	–	11.1	88.5
Omnidata [3]	11.9M	310K	7.4	94.5
DPT [12]	1.2M	188K	9.8	90.3
Painter [20]	24K	–	8.0	95.0
Marigold [6]	–	74K	5.5	96.4
DMP	–	10K	12.0	86.5

ADE20K. We also investigate the performance of semantic segmentation with more classes, *e.g.* 150 classes in ADE20K [22]. We follow the encoding strategy proposed by Wang et al. [20] and convert class indices into 3-digit numbers with a b-base system, which can be represented in the RGB space. However, the performance is unsatisfying (lower than 20% accuracy). With the number of classes increasing, the differences between colors are less distin-

Table 2. Comparison of performance between the subset of ADE20K [22] and the synthesized bedrooms.

		Bed	Pillow	Lamp	Window	Painting	Mean
ADE20K	Acc↑	0.88	0.36	0.57	0.76	0.74	0.66
	mIoU↑	0.82	0.25	0.39	0.60	0.60	0.53
Bedrooms	Acc↑	0.89	0.59	0.64	0.83	0.75	0.75
	mIoU↑	0.85	0.36	0.44	0.73	0.67	0.61

Table 3. **Analysis of training cross-attention layers and providing text condition.** Both improve the performance of in-domain samples but make little difference in out-of-domain data.

	In-domain		Out-of-domain	
	L1↓	Ang↓	L1↓	Ang↓
Self-attn	0.0606	0.1290	0.0890	0.1871
Self-attn + text	0.0605	0.1293	0.0876	**0.1844**
All attn + text	**0.0514**	**0.1156**	**0.0872**	0.1886

guishable. The unlabeled areas, which can be simply ignored when calculating loss in the image space, become hard to tackle in the latent space. We leave the application of real-world many-class semantic segmentation for future exploration.

We conduct another analysis with a subset of ADE20K containing only images with beds. The train-test split is constructed by applying the same filtering to the original splits, resulting in 1825 training images and 189 test images. We also generate prompts with BLIP-2. The results are presented in Table 2. The performance across real and synthetic domains is similar, especially for large items.

B.3. Additional Ablation Study

Modeling. We vary the trainable layers and the presence of text conditions when fine-tuning the model. Since the example tasks we choose in this work are not directly conditional on text, providing text descriptions might not be necessary. Accordingly, training cross-attention layers is optional. Table 3 shows that text condition and training cross-attention layers help improve the performance of in-domain samples, but the difference in out-of-domain samples is unnoticeable between the settings. This result suggests that we can adopt curated real ground truth datasets without text descriptions for training at the expense of a subtle performance drop. Alternatively, we can generate prompts with image captioning [7], which may lead to better performance.

Size of Training Data. We analyze the effect of varying the size of training data. We compare fine-tuning the model for normal estimation with 100, 1K, 10K, and 100K generated bedroom images. As shown in Figure 2, increasing

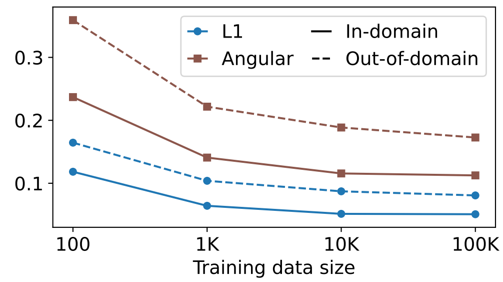


Figure 2. Quantitative performance of normal estimation with different sizes of training data.

Table 4. NYU Depth v2 [18] performance comparison of models trained with real and pseudo ground truths.

Dataset	Ground Truth	REL↓	$\delta\uparrow$
Hypersim [14]	Real	13.0	85.0
Bedrooms	Pseudo	12.0	86.5

data size over 10K improves little performance, so we conduct the other experiments with 10K training images.

Quality of Training Data. We examine the influence of data quality by comparing the models trained with real and pseudo ground truth. We use Hypersim [14] as the real ground truth and evaluate the models with NYU Depth v2 [18]. Table 4 shows that there is no significant difference between the two models. The model trained with pseudo ground truth even performs slightly better. We speculate that the data diversity may be an important factor. While Hypersim contains more than 70K images, the images are collected from only 461 scenes. Many of them are variations of camera views and distances. In contrast, the synthetic images, while all of them are bedrooms, are all distinct scenes, which present diverse compositions of objects.

Blending Inputs and Outputs. IADB [4] proposes a deterministic framework where the diffusion process is formulated as a series of interpolations between observations and noise. Although their training strategy produces deterministic mapping of observations and noise, the correlation between observation and noise in each pair is stochastic due to unpaired sampling during training. We analyze the applicability of this framework to deterministic dense prediction problems by sampling paired inputs and outputs and fine-tuning from a pre-trained T2I diffusion model. With such adaptation, the differences between their framework and our approach are only the variance schedule and parametrization, where the importance weight of inputs linearly rises through the diffusion process, and the U-Net predicts $y - x$.

Table 5 shows the comparison between `DMP` and IADB

Table 5. Comparions with IADB [4] and Poission blending [11] on surface normal estimation.

	In-domain		Out-of-domain	
	L1↓	Ang↓	L1↓	Ang↓
IADB [4]	0.0675	0.1416	0.0974	0.2017
Poission [11]	0.0868	0.1888	0.1201	0.2623
DMP	**0.0514**	**0.1156**	**0.0872**	**0.1886**

Table 6. Comparions with IADB [4] on depth estimation.

	In-domain			Out-of-domain		
	REL↓	$\delta\uparrow$	RMSE↓	REL↓	$\delta\uparrow$	RMSE↓
IADB [4]	0.3099	0.4982	0.1165	0.5049	0.3132	0.1467
DMP	**0.1072**	**0.8861**	**0.1020**	**0.2117**	**0.6395**	**0.1360**

on surface normal estimation, and Table 6 is the result of depth estimation. Figure 3 demonstrates that the images generated by the model fine-tuned through IADB have noise and inaccurate predictions.

In addition to α-blending used by DMP and IADB, we investigate the effect of an advanced blending strategy, Poission blending [11], which blends source and target images by solving a least-square fitting while reserving the gradient of source images. We assume image gradients are meaningful in the latent space. The diffusion process is viewed as increasing the intensity of the mask for selection editing. We adopt an off-the-shelf PyTorch implementation [2]. The performance on surface normal estimation is shown in Table 5, and the example outputs in Figure 3 show that the image quality is unsatisfying.

B.4. Additional Comparison

ControlNet. ControlNet [21] proposes a conditional text-to-image framework with additional control, such as edges or human poses, which constrains structures and layouts of output images. Since it is also an image-to-image generative model, we train it to take input images as control and output estimations. The performance of estimating 3D properties and intrinsic images is presented in Table 7, and the segmentation results are shown in Table 8. It demonstrates weaker generalizability than our approach.

In addition, we analyze the influence of varying initial noise in Figure 4. While the rough structures of the images are controlled by the input images, the initial noise alters the details of estimations. This variation is not tolerated for dense prediction.

Palette. Besides training GAN-based generative models from scratch and fine-tuning pre-trained diffusion models with the approaches listed in Section 4.1, we additionally include training an image-to-image diffusion model from scratch for comparison. Following the design of Palette [15], we expand the input layers of the U-Net to encode the concatenation of input and output images, with the U-Net parameterized to predict noise. The same autoencoder in the pre-trained diffusion model is also adopted. The performance is shown in Table 7 and Table 8, which indicates the inability of this approach to handle categorical label maps.

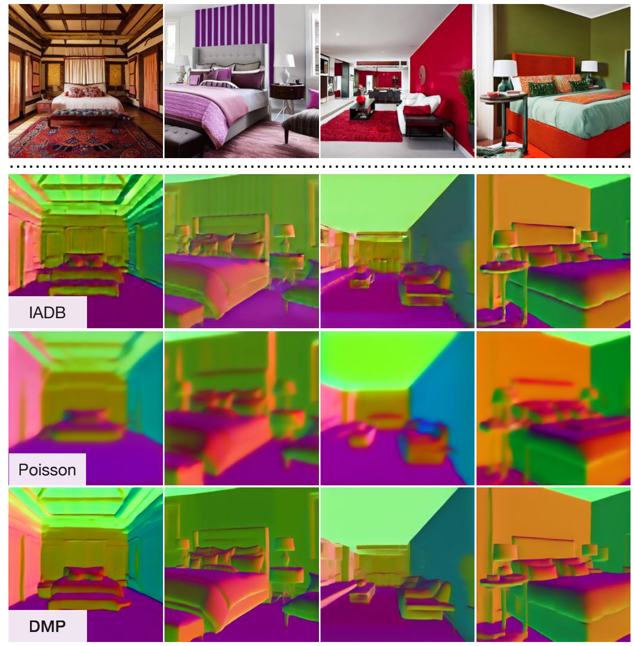


Figure 3. Qualitative comparisons between different blending frameworks.

Figure 4. **Results of ControlNet with different initial noise.** The outputs are not deterministic.

Table 7. Quantitative comparisons with ControlNet [21] and Palette [15] on 3D property estimation and intrinsic image decomposition.

	Normal				Depth						Albedo		Shading	
	In		Out		In			Out			In	Out	In	Out
	L1↓	Ang↓	L1↓	Ang↓	REL↓	$\delta\uparrow$	RMSE↓	REL↓	$\delta\uparrow$	RMSE↓	MSE↓	MSE↓	MSE↓	MSE↓
ControlNet [21]	0.1021	0.2216	0.1862	0.4032	0.1739	0.7681	0.1287	0.4398	0.4004	0.2253	0.0302	0.0402	0.0265	0.0336
Palette [15]	0.1643	0.3642	0.1881	0.4160	0.6889	0.2626	0.3604	1.0535	0.2270	0.4203	0.0203	0.0199	0.0304	0.0260
DMP	**0.0514**	**0.1156**	**0.0872**	**0.1886**	**0.1072**	**0.8861**	**0.0041**	**0.2117**	**0.6395**	**0.1360**	**0.0051**	**0.0064**	**0.1020**	**0.0070**

Table 8. Quantitative comparisons with ControlNet [21] and Palette [15] on semantic segmentation.

	Bed		Pillow		Lamp		Window		Painting		Mean	
	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑	Acc↑	mIoU↑
ControlNet [21]	0.5215	0.4820	0.3540	0.1436	0.4275	0.2936	0.4999	0.4190	0.3823	0.3257	0.4370	0.3328
Palette [15]	0.0347	0.0329	0.0019	0.0018	0.0013	0.0012	0.0119	0.0119	0.0005	0.0005	0.0101	0.0097
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**

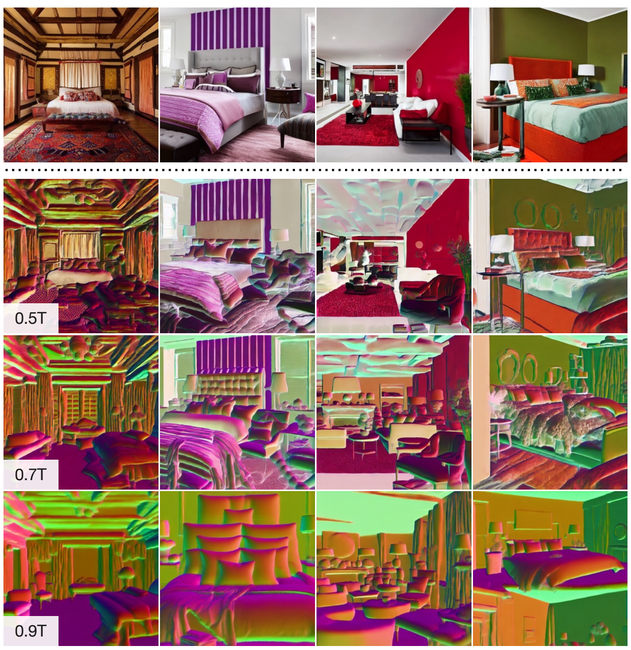


Figure 5. **Qualitative comparisons of SDEdit with starting from different time steps.** A trade-off exists between the effect of style transfer and content preservation.

B.5. Improving Compared Methods

Since the results of GAN-based generative models consistently outperform diffusion-based models in our experiments, we seek performance enhancement for diffusion-based approaches. All experiments are conducted on in-domain surface normal estimation.

Table 9. **Quantitative comparisons on in-domain surface normal estimation between different timesteps where the generation process of SDEdit starts.** The performance improves at the expense of deviation from input image contents.

Step	L1↓	Ang↓
$0.5T$	0.2897	0.5336
$0.7T$	0.2599	0.5087
$0.9T$	0.2059	0.4568

Table 10. **Quantitative comparisons on in-domain surface normal estimation between DDIB and DDIB with Plug-and-Play (PnP).** The feature injection regulates the generated contents while improving performance.

Variants	L1↓	Ang↓
DDIB	0.1849	0.4210
DDIB + PnP	0.1652	0.3634

SDEdit. The time steps from which the generation process of SDEdit starts can be seen as the strength of preserving the contents of input images. We show in Figure 5 that generating from step $0.5T$ produces images with similar contents to the input images, while from step $0.9T$ results in plausible estimation of surface normals, but the image contents are disrupted, despite achieving the best performance in quantitative evaluation reported in Table 9. This issue has long been understood as a trade-off between the effect of style transfer and content preservation in image-to-image literature [5, 8], but for deterministic dense prediction problems considered in this work, such a trade-off is not permitted.

Figure 6. **Qualitative comparisons between DDIB and DDIB with Plug-and-Play (PnP).** The image contents are reserved but not consistent with accurate normals.

Table 11. **Quantitative comparisons on in-domain surface normal estimation between different training tokens of IP2P (learned).** The increased number of tokens does not guarantee improved performance.

#Tokens	L1↓	Ang↓
1	0.3550	0.7181
2	0.3470	0.7790
4	0.3274	0.6384

DDIB. As presented in Appendix B.7, DDIB is capable of generating images that are likely sampled from output distributions, but the contents and geometry of output results are not consistent with input images. We explore the approach to content consistency by adopting feature constraints proposed by Plug-and-Play (PnP) [19] for image-to-image translation, which injects the self-attention and convolution features of input images into output images. As shown in Figure 6 and Table 10, the structures and contents of output images of DDIB with PnP constraints highly resemble the input images, but the estimated normals remain inaccurate despite better quantitative performance.

IP2P. We analyze the expressiveness of inverted tokens by varying the number of training tokens in IP2P (learned). While the performance is slightly improved in one metric of quantitative evaluation, shown in Table 11, Figure 7 reveals that the differences between the estimated results are not significant.

B.6. Failure Cases

We demonstrate some examples of failure cases in Figure 8 for surface normal estimation, Figure 9 for depth estimation, and Figure 10 for semantic segmentation, where off-the-shelf approaches might provide more accurate prediction than our method.

Figure 7. Qualitative comparisons of IP2P with different training tokens.

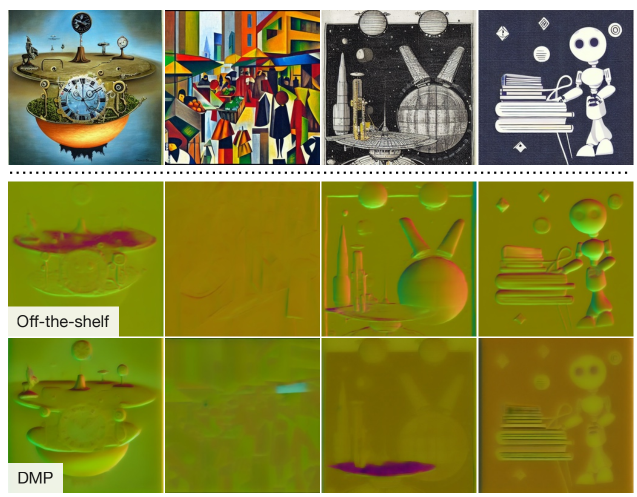


Figure 8. Failure cases of surface normal estimation.

B.7. Results of Compared Methods

We show the example images generated by the compared methods listed in Section 4.1. The results of surface normal estimation are in Figure 11, with depths in Figure 12, albedo in Figure 13, shading in Figure 14, and semantic segmentation in Figure 15.

Table 12. **Style templates**, where {} is replaced by original prompts.

- anime artwork, {} . anime style, key visual, vibrant, studio anime, highly detailed
- concept art, {} . digital artwork, illustrative, painterly, matte painting, highly detailed
- comic, {} . graphic illustration, comic art, graphic novel art, vibrant, highly detailed
- neonpunk style, {} . cyberpunk, vaporwave, neon, vibes, vibrant, stunningly beautiful, crisp, detailed, sleek, ultramodern, magenta highlights, dark purple shadows, high contrast, cinematic, ultra-detailed, intricate, professional
  surrealist art, {} . dreamlike, mysterious, provocative, symbolic, intricate, detailed
- abstract style, {} . non-representational, colors and shapes, expression of feelings, imaginative, highly detailed
- art deco style, {} . geometric shapes, bold colors, luxurious, elegant, decorative, symmetrical, ornate, detailed
- vaporwave style, {} . retro aesthetic, cyberpunk, vibrant, neon colors, vintage 80s and 90s style, highly detailed

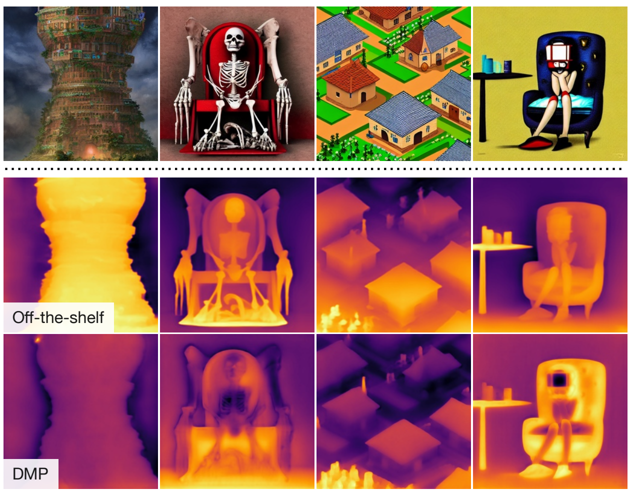


Figure 9. Failure cases of depth estimation.

Figure 10. Failure cases of semantic segmentation.

C. Implementation Details

Model Architecture and Optimization. We use Stable Diffusion 1.4 as the pre-trained text-to-image model and adapt it with rank $= 4$ for LoRA. We fine-tune the model for 50K steps with batch size 8 and learning rate 0.0001 with a cosine decay schedule. The training takes around 14 hours with a single NVIDIA RTX 3090.

Generating Images. We generate the training and test images by first generating a set of prompts with a large language model. The prompt for the language model is a template adapted from pix2pix-zero [10], where different scene keywords are filled in. The template is

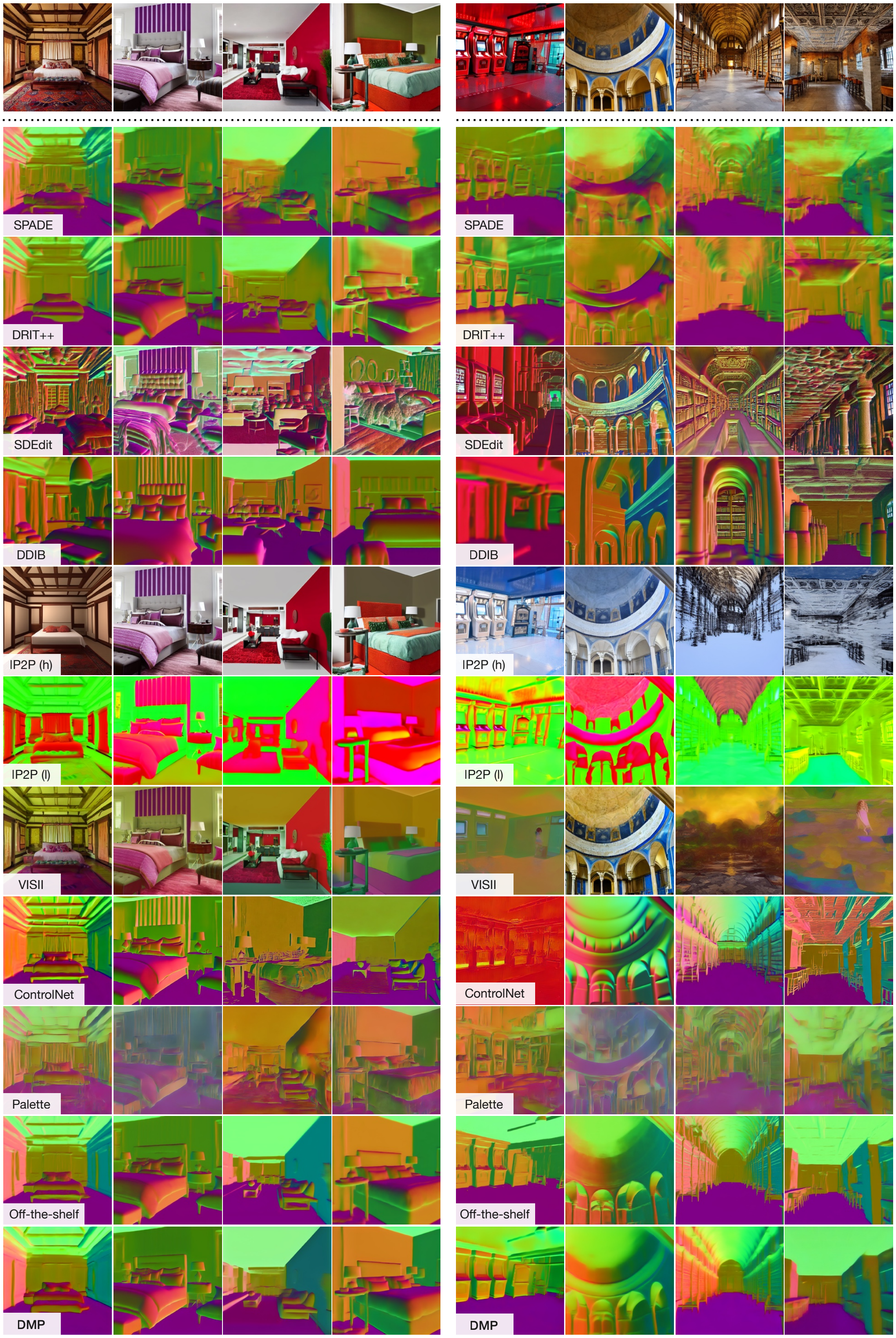
> "Provide a caption for a photo of a `scene`. The caption should contain many adjectives, should describe colors, styles, lighting and materials in the photo, should be in English and should be no longer than 150 characters. Caption:".

The placeholder `scene` is replaced by "bedroom" for training images and in-domain test images. To generate out-of-domain test images for estimating 3D properties and intrinsic images, it is replaced by uniform sampling from the keywords in Table 13.

Out-of-domain test images for segmentation are synthesized by varying the image styles of in-domain test images, for semantic categories should remain the same across training and test images. The prompts regulating the styles are listed in Table 12 borrowed from an online post [1].

D. Applications

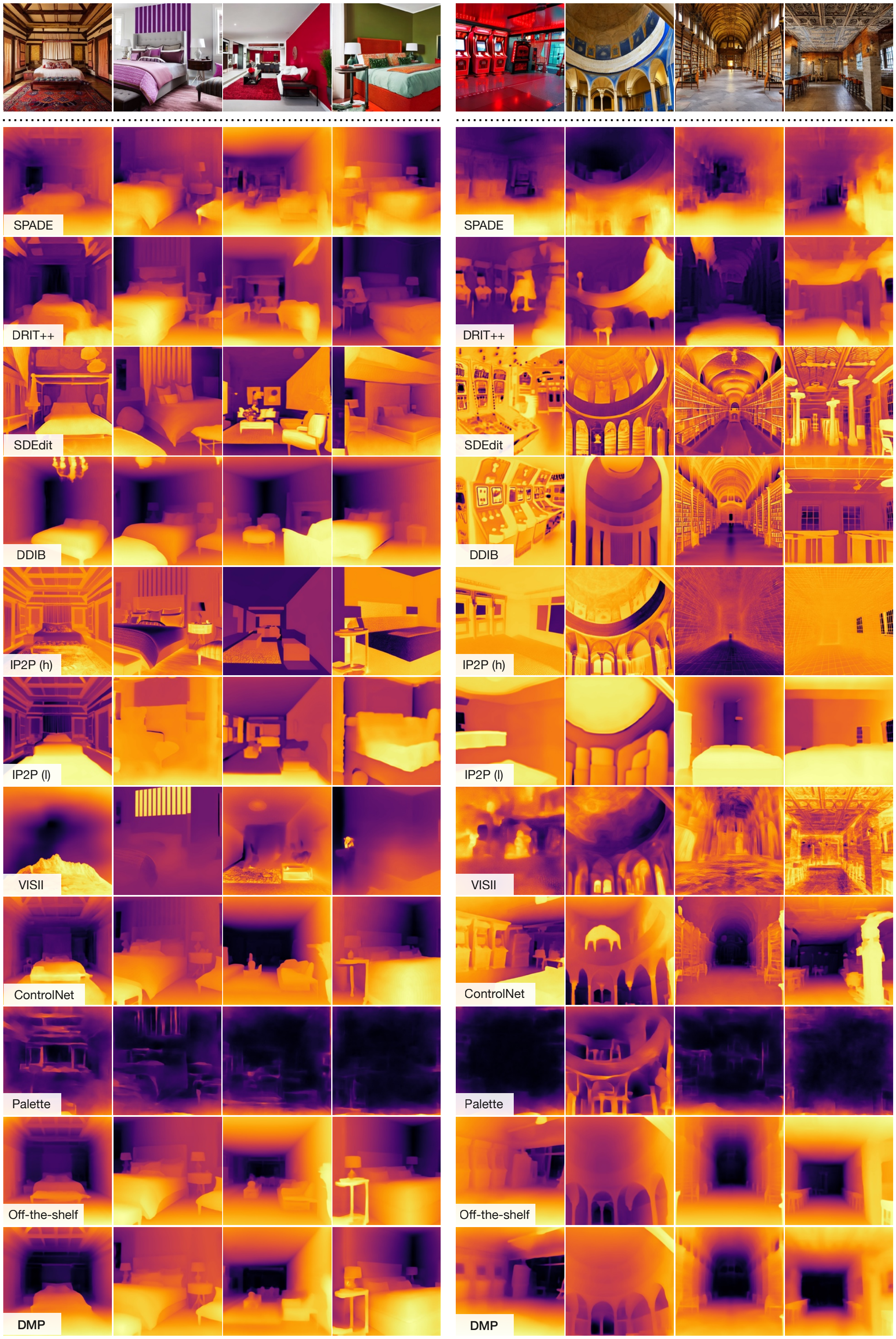
Surface normals and depths facilitate many vision tasks. We show by the examples of 3D photo inpainting [17] that precise depths improve 3D reconstruction from 2D images. Compared to the default depth estimator [13], the resulting videos produced with the depth maps generated by our approach have more accurate depth relationships between the objects. Please refer to the project website for visual demonstrations.



(a) In-domain

(b) Out-of-domain

Figure 11. Qualitative results of compared methods on surface normal estimation.



(a) In-domain

(b) Out-of-domain

Figure 12. Qualitative results of compared methods on depth estimation.

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

Figure 13. Qualitative results of compared methods on albedo estimation.

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

Figure 14. Qualitative results of compared methods on shading estimation.

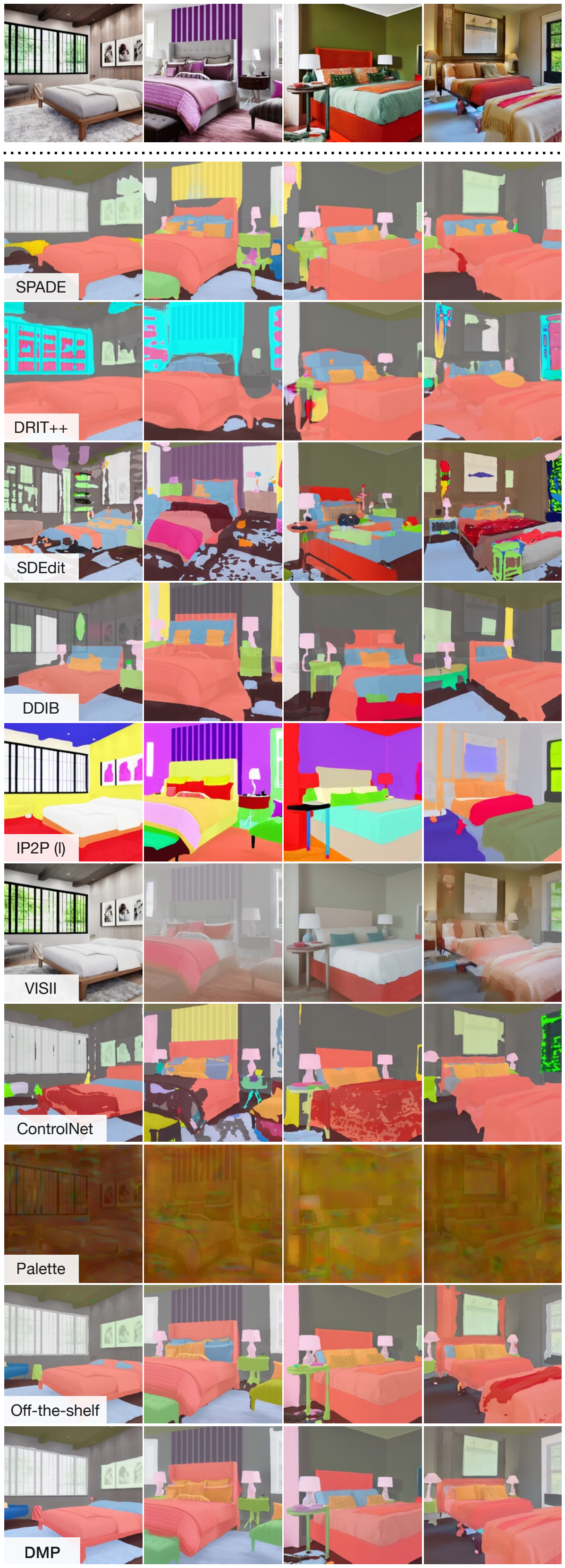


Figure 15. Qualitative results of compared methods on semantic segmentation.

Table 13. **Scenes categories** of out-of-domain images.

airlock	airplane cabin	airport terminal	airport ticket counter
alcove	amusement arcade	anechoic chamber	indoor apse
aquarium	arcade	archive	armory
indoor arrival gate	art gallery	art school	art studio
artists loft	assembly line	indoor athletic field	attic
auditorium	auto factory	indoor auto mechanics	auto showroom
backstage	indoor badminton court	baggage claim	ball pit
ballroom	indoor bank	bank vault	banquet hall
indoor baptistry	bar	barbershop	barrack
basement	indoor basketball court	bathhouse	bathroom
indoor batting cage	indoor bazaar	beauty salon	bedchamber
bedroom	beer hall	belfry	bell foundry
berth	berth deck	betting shop	bicycle racks
bindery	biology laboratory	indoor bistro	indoor bleachers
indoor bomb shelter	bookbindery	bookstore	indoor booth
indoor bow window	bowling alley	box seat	boxing ring
breakroom	indoor brewery	indoor brickyard	burial chamber
indoor bus depot	bus interior	indoor bus station	butchers shop
indoor cabin	cafeteria	call center	candy store
canteen	backseat car interior	frontseat car interior	cardroom
cargo container interior	indoor carport	indoor casino	catacomb
indoor cathedral	catwalk	chapel	checkout counter
cheese factory	chemistry lab	indoor chicken coop	indoor chicken farm
childs room	interior choir loft	indoor church	indoor circus tent
classroom	clean room	indoor clock tower	indoor cloister
closet	clothing store	cockpit	coffee shop
computer room	conference center	conference hall	conference room
confessional	control room	indoor control tower	indoor convenience store
corridor	courtroom	interior covered bridge	crawl space
cybercafe	indoor dairy	dance school	darkroom
day care center	delicatessen	dentists office	department store
departure lounge	indoor diner	dining car	dining hall
dining room	discotheque	distillery	indoor doorway
dorm room	dress shop	dressing room	indoor driving range
drugstore	editing room	electrical room	elevated catwalk
interior elevator	elevator lobby	elevator shaft	engine room
entrance hall	indoor escalator	exhibition hall	fabric store
indoor factory	fastfood restaurant	indoor ferryboat	indoor firing range
fishmarket	interior fitting room	indoor flea market	indoor florist shop
food court	indoor foundry	funeral chapel	funeral home
furnace room	galley	game room	indoor garage
indoor general store	indoor geodesic dome	gift shop	great hall
indoor greenhouse	indoor gun deck	gun store	indoor gymnasium
hallway	indoor hangar	hardware store	hat shop
hatchery	hatchway	hayloft	hearth
home office	home theater	hospital room	indoor hot tub
hotel breakfast area	hotel room	indoor hunting lodge	ice cream parlor
indoor ice skating rink	indoor incinerator	indoor inn	indoor jacuzzi
indoor jail	jail cell	jewelry shop	jury box
indoor kennel	kindergarden classroom	indoor kiosk	kitchen
kitchenette	lab classroom	indoor labyrinth	landing

laundromat	lavatory	lecture room	legislative chamber
indoor library	indoor lido deck	limousine interior	indoor liquor store
living room	lobby	locker room	loft
indoor lookout station	indoor lumberyard	machine shop	indoor market
martial arts gym	maternity ward	mess hall	mezzanine
military hospital	mill	mine	indoor mini golf course
indoor monastery	morgue	indoor mosque	indoor movie theater
indoor museum	music store	music studio	natural history museum
newsroom	indoor newsstand	nightclub	indoor nuclear power plant
nursery	nursing home	indoor observatory	office
office cubicles	indoor oil refinery	operating room	optician
orchestra pit	interior organ loft	orlop deck	ossuary
indoor outhouse	oyster bar	packaging plant	palace hall
pantry	paper mill	indoor parking garage	parlor
particle accelerator	indoor party tent	pawnshop	penalty box
perfume shop	pet shop	pharmacy	physics laboratory
piano store	pig farm	indoor pilothouse	pizzeria
indoor planetarium	playroom	indoor podium	portrait studio
indoor power plant	print shop	promenade deck	indoor pub
pulpit	pump room	indoor quonset hut	reading room
reception	recreation room	indoor recycling plant	refectory
repair shop	restaurant	restaurant kitchen	indoor restroom
revolving door	riding arena	indoor roller skating rink	rolling mill
sacristy	sauna	sawmill	science museum
scriptorium	security check point	server room	sewer
sewing room	shipping room	indoor shipyard	shoe shop
indoor shopping mall	shower	shower room	shrine
indoor skywalk	sporting goods store	squash court	stable
indoor stage	staircase	indoor steam plant	indoor steel mill
storage room	storeroom food	submarine interior	subway interior
supermarket	sushi bar	indoor swimming pool	indoor synagogue
tearoom	teashop	television room	television studio
indoor tennis court	indoor tent	textile mill	indoor procenium theater
indoor round theater	indoor seats theater	thriftshop	throne room
ticket booth	indoor ticket window	indoor tobacco shop	toyshop
indoor track	trading floor	train interior	rail indoor tunnel
road indoor tunnel	turkish bath	utility room	utility tunnel
van interior	indoor velodrome	ventilation shaft	vestry
veterinarians office	videostore	indoor volleyball court	voting booth
waiting room	walk in freezer	indoor warehouse	indoor washhouse
indoor water treatment plant	wet bar	whispering gallery	wig shop
window seat	winery	witness stand	workroom
workshop	indoor wrestling ring	youth hostel	basketball arena
football arena	hockey arena	performance arena	rodeo arena
soccer arena	home atrium	public atrium	bakery kitchen
bakery shop	airplane cargo deck	choir loft	cloakroom booth
cloakroom	library cubicle	office cubicle	home dinette
vehicle dinette	elevator door	freight elevator	ferryboat cargo deck
fitting room	organ loft	establishment poolroom	home poolroom
spa massage room	spa mineral bath	corridor in a subway station	platform in a subway station
turnstiles in a subway station	platform in a train station	station in a train station	barrel storage in a wine cellar
bottle storage in a wine cellar			

References

[1] Andrew. 106 styles for stable diffusion xl model. https://stable-diffusion-art.com/sdxl-styles. 6

[2] Matthew Baugh. PIE-torch. https://github.com/matt-baugh/pytorch-poisson-image-editing. 3

[3] 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. 1

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

[5] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017. 4

[6] 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*, 2024. 1

[7] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 1, 2

[8] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. In *NeurIPS*, 2017. 4

[9] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019. 1

[10] 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. 6

[11] Patrick Pérez, Michel Gangnet, and Andrew Blake. Poisson image editing. *ACM TOG*, 2003. 3

[12] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 1

[13] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44 (03):1623–1637, 2022. 1, 6

[14] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *ICCV*, 2021. 2

[15] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH*, 2022. 3, 4

[16] Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *ICLR*, 2022. 1

[17] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *CVPR*, 2020. 6

[18] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 1, 2

[19] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *CVPR*, 2023. 5

[20] 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. 1

[21] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 3, 4

[22] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 1, 2