Self-Predicted Depth-Aware Guidance with Diffusion Models

Gyeongnyeon Kim*, Wooseok Jang*, Gyuseong Lee*, Susung Hong, Junyoung Seo, Seungryong Kim†
Korea University, Seoul, Korea
{kkn9975,jws1997,jpl358,susung1999,se780,seungryong_kim}@korea.ac.kr

Abstract

Generative models have recently undergone significant advancement thanks to diffusion models, which use effective guidance techniques like classifier or classifier-free guidance to trade-off fidelity and diversity. However, these methods are not capable of guiding a generated image to be aware of its geometric configuration, e.g., depth, which hinders their application to areas that require a certain level of depth awareness. To address this limitation, we propose a novel guidance method for diffusion models that uses self-estimated depth information derived from the rich intermediate representations of diffusion models. Concretely, we first present label-efficient depth estimation framework using internal representations of diffusion models. Subsequently, we propose the incorporation of two guidance techniques based on pseudo-labeling and depth-domain diffusion prior during the sampling phase to self-condition the generated image using the estimated depth map. Our experiments show that our method effectively guides diffusion models to generate geometrically plausible images.

1. Introduction

Diffusion models [10, 16, 19, 24] have recently received much attention and have shown remarkable generation quality and diversity. However, those works hardly consider geometrical configuration during the image generation process. As a result, conventional diffusion models often generate geometrically implausible images that contain ambiguous depth and cluttered object layouts which can be visually unappealing but also unsuitable for downstream tasks, e.g., robotics and autonomous driving [28, 30].

While some approaches [7, 11] drive the sampling process of diffusion models toward a class-specific distribution, limited attention has been given to guiding diffusion models towards a geometrically plausible image distribution. To address this, we propose Depth-Aware Guidance (DAG), which incorporates depth awareness into diffusion models. Training both diffusion models and depth predictors from scratch is challenging, so we use a pretrained diffusion model's rich representations to train depth predictors, extending our knowledge of diffusion models' representation capabilities in depth prediction tasks.

Furthermore, by leveraging the label-efficient depth predictors, we propose two depth-aware guidance strategies for geometric awareness: depth consistency guidance and depth prior guidance. Depth consistency guidance (DCG) uses consistency regularization [2, 12, 21] to guide the image towards improving the poor prediction by treating the better prediction as a depth pseudo-label. Depth prior guidance (DPG) utilizes an additional pretrained diffusion U-Net as a prior network [9, 17] to provide guidance during the sampling process and explicitly injects depth information into the sampling process of diffusion models.

To evaluate our framework, we conduct experiments on indoor and outdoor scene datasets [15, 31] and propose new metric from the perspective of depth estimation tasks to capture geometric awareness of the generated images. To the best of our knowledge, our work is the first attempt to utilize depth information during the sampling process to make image generation more aware of geometric configuration.

2. Methodology

2.1. Label-Efficient Training of Depth Predictors

In order to generate depth-aware images with diffusion guidance in a straightforward way as in [7], we need either a large amount of image-depth pairs or an external large-scale depth estimation network trained on the noised images, both of which are challenging to acquire. To address this problem, we propose to re-use the rich representations learned with DDPM [10] that contain depth information of images to estimate the depth.

Network architecture. Recent research has shown that the internal features of the networks trained with diffusion models can encode semantic information [1, 4, 29], and our contribution builds upon this by incorporating depth information into the framework.

*Authors contributed equally.
†Corresponding author.

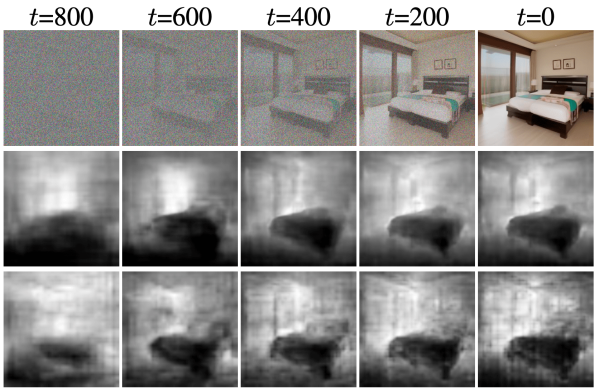


Figure 1. **Visualizations of the sampling process of our framework:** (from top to bottom) predicted images, depth predictions from the strong-branch predictor, and depth predictions from the weak-branch predictor. As exemplified, the strong-branch predictor gives robust depth predictions even at the early stage.

For label-efficient depth estimation, we utilize a pixel-wise shallow MLP regressor. Specifically, we acquire the internal features $\mathbf{f}_t(k) \in \mathbb{R}^{C(k)\times H(k)\times W(k)}$ from the output of k-th decoder layer in the diffusion U-Net, where $C(k)$ denotes the channel dimension and $H(k) \times W(k)$ denotes the spatial resolution of the k-th layer of the U-Net decoder. Then, we form the depth map by querying the MLP blocks pixel-by-pixel, where the depth map can be formulated as:

$$\mathbf{d}_t(k) = \text{MLP}(\mathbf{f}_t(k)). \tag{1}$$

In this setting, it may be better to use more features from different U-Net layers than only using the feature from one layer [4]. Hence, we extract features from multiple layers and concatenate them in a channel dimension to obtain $\mathbf{g}_t = [\mathbf{f}_t(1), \mathbf{f}_t(2), \cdots, \mathbf{f}_t(d)]$, where d is the total number of selected layers. These features are then passed to the pixel-wise MLP depth predictor. Additionally, we append a time-embedding block similar to the diffusion U-Net's time-embedding module to the depth predictor input, enabling prediction at any timestep and throughout the sampling process. Applying it to Eq. 1, we can achieve

$$\mathbf{d}_t = \text{MLP}(\mathbf{g}_t, t). \tag{2}$$

Loss function. We train the depth estimator only with the frozen features from the diffusion U-Net by using the ground-truth depth map $\mathbf{y}$ with L1 loss as

$$\mathcal{L}_{\text{depth}} = \|\mathbf{d}_t - \mathbf{y}\|_1. \tag{3}$$

This whole procedure allows us to achieve reasonable label-efficient prediction performance in the depth domain, as demonstrated in Fig. 2(b). The depth prediction scheme allows us to predict the depth map for the intermediate images under generation in arbitrary sampling steps, as shown in Fig. 1, since the representations of diffusion models are inherently learned with the timesteps.

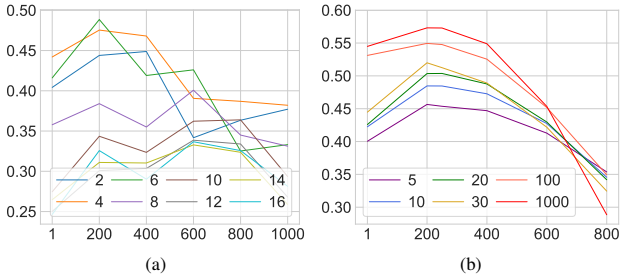


Figure 2. **Quantitative comparisons of depth prediction performance.** Evaluation of depth estimation performance for varying timesteps with (a) different U-Net blocks and (b) training image numbers. The x-axis shows timesteps and the y-axis shows depth estimation accuracy.

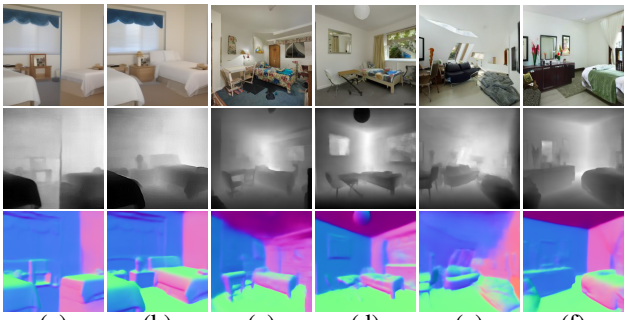
Figure 3. **Visualization of depth estimation on the ImageNet dataset [6].** The images in the upper row are from ImageNet, and images in the middle row are estimated depth maps using our depth estimator which takes the inner features from the U-Net of the Latent Diffusion [19] as input. The prediction is done with noisy images.

2.2. Depth Guided Sampling for Diffusion Model

To ensure plausible depth maps from generated images, we encourage the predicted depth maps to be accurate during sampling. To this end, we propose two guidance techniques that use utilize the aforementioned efficiently-trained depth predictors in Sec. 2.1. Because of the absence of a pre-determined label, we cannot naively compute the loss for guiding the sampling process. Therefore we build two alternative loss functions that act as guidance constraints. We discuss the details in the following sections. The general form of guidance equation is formulated by:

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t) - \omega\nabla_{\mathbf{x}_t}\mathcal{L}_{\text{depth}}, \Sigma_\theta(\mathbf{x}_t)). \tag{4}$$

Depth consistency guidance. Pseudo-labeling [13, 25] can a be a possible approach for depth estimation in the absence of ground truth information, but generating confident predictions for pseudo-labeling is challenging. Inspired by FixMatch [25], our method combines pseudo-labeling [13] with consistency regularization [2, 21], using weakly-augmented labels as pseudo-labels to enhance performance. We propose that richer representations lead to more accurate depth maps, and thus we consider predictions from multiple feature blocks as *strong branch* predictions



(a) (b) (c) (d) (e) (f)

Figure 4. **Qualitative comparison on LSUN-bedroom [31].** We visualize (top) the samples without guidance ((a), (c), (e)) and with depth-aware guidance ((b), (d), (f)), and their corresponding depths [18] (middle) and surface normals [3] (bottom).

and those from fewer features as *weak branch* predictions. These strong branch predictions are more informative and suitable for use as robust pseudo-labels (Fig. 1).

In specific, we use the features $\mathbf{g}^W = [\mathbf{f}_t(6)]$ for weak branch features, and $\mathbf{g}^S = [\mathbf{f}_t(2), \mathbf{f}_t(4), \mathbf{f}_t(5)\mathbf{f}_t(6), \mathbf{f}_t(7)]$ for strong branch features. To account for the different channel dimensions of these two aggregated features, we design two asymmetric predictors: MLP-S and MLP-W. The first predictor receives more features from the U-Net block, while the second one receives fewer features, and we train them together. We feed the collected features to these MLPs and obtain the depth map predictions:

$$\mathbf{d}_t^S = \text{MLP-S}(\mathbf{g}_t^S, t), \quad \mathbf{d}_t^W = \text{MLP-W}(\mathbf{g}_t^W, t). \tag{5}$$

As stated above, we treat $\mathbf{d}_t^S$ as a pseudo-label and $\mathbf{d}_t^W$ as a prediction then compute the loss using a consistency loss between two predicted dense maps. We apply the stop-gradient operation to the strong features, preventing the strong prediction from learning the weak prediction [5, 25]. The gradient of the loss with respect to $\mathbf{x}$ flows through the diffusion U-Net and guides the sampling process as done in [7]. This process can be formulated as

$$\mathcal{L}_{\text{dc}} = \|\text{stopgrad}(\mathbf{d}_t^S) - \mathbf{d}_t^W\|_2^2, \tag{6}$$

where stopgrad denotes the stop-gradient operation.

Depth prior guidance. We also propose another guidance method, which we call depth prior guidance, to inject depth prior into the sampling process. The pretrained diffusion model can effectively refine the noised distributions to realistic distributions [14], or it can help to optimize the noised initialization of the data to match with the real data by utilizing the knowledge of the diffusion model [9, 17]. Therefore we train another small-resolution diffusion U-Net ϵ_ϕ on the depth domain and use it as our prior for the second guidance method. As described in Sec. 2.1, we can extract

Figure 5. **Visualization of point cloud representation obtained by depth information.** We compare the results generated from the baseline without (odd rows) and with (even rows) our guidance by showing images and transforming them into point cloud visualizations in four different views.

the features from the decoder part of the image-generating U-Net to estimate the corresponding depth map using MLP depth predictor. Then, we inject noise to the depth prediction $\mathbf{d}_0^S$ using a forward process of diffusion like:

$$\mathbf{d}_\tau^S = \sqrt{\bar{\alpha}_\tau}\mathbf{d}_0^S + \sqrt{1-\bar{\alpha}_\tau}\eta, \quad \eta \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{7}$$

where τ is the timestep that is used in a prior diffusion model. After adding noise to the depth prediction, we feed it to our prior network to estimate the added noise. Then we calculate the gradient of the mean-squared error between the added noise and the predicted noise concerning x. This process is then defined as:

$$\mathcal{L}_{\text{dp}} = \|\eta - \epsilon_\phi(\mathbf{d}_\tau^S))\|_2^2. \tag{8}$$

Overall guidance. To integrate the proposed DCG and DPG, we calculate the gradients of Eq. 6 and 8. As in [7], our overall sampling can be written as:

$$\mathbf{x}_{t-1} \sim \mathcal{N}(\mu_\theta(\mathbf{x}_t) - \omega_{\text{dc}}\nabla_{\mathbf{x}_t}\mathcal{L}_{\text{dc}} - \omega_{\text{dp}}\nabla_{\mathbf{x}_t}\mathcal{L}_{\text{dp}}, \Sigma_\theta(\mathbf{x}_t)). \tag{9}$$

where ω_{dc} and ω_{dp} denotes the DCG scale and DPG scale respectively.

3. Experiments

3.1. Experimental Settings

To evaluate the performance of our proposed method, we conduct experiments on LSUN-Bedroom and LSUN-Church [31] for both depth estimation and image generation tasks. As there are no ground-truth depth labels available in the LSUN dataset, we generate pseudo-labels using

Methods	DPG	DCG	dFID (↓)
Baseline	-	-	15.71
DAG	✓	-	14.18
	-	✓	15.27
	✓	✓	**13.93**

Table 1. **Quantitative Results on the LSUN-bedroom [31].**

Methods	DPG	DCG	dFID (↓)
Baseline	-	-	17.69
DAG	✓	-	17.43
	-	✓	17.40
	✓	✓	**17.31**

Table 2. **Quantitative results on the LSUN-church [31].**

a DPT [18] pretrained on the NYU-Depth dataset [23] and utilize them for training the depth estimator.

We incorporate geometric awareness in the image generation process and introduce a novel performance metric for models that generate depth-guided images. First, we predict the depth maps of generated images using Dense Prediction Transformer (DPT) [18]. To measure the reality of the depth estimation map, we directly evaluate FID [22] with depth images and denote it dFID. To make a fair comparison, we build the reference batch following [7] with the depth predictions of images from the dataset with DPT-Hybrid.

3.2. Experimental Results

Depth prediction performance. First, to provide guidance for synthesizing images, we train the predictor for timesteps $t < 800$. We evaluate the depth performance using the depth accuracy metric ($\delta < 1.25$), and the results are shown in Fig. 2. Based on the results in Fig. 2 (a), we choose to use the middle feature blocks $\{l_n\} = \{2, 4, 5, 6, 7\}$, which show relatively high accuracy. In DCG, we also choose the feature maps by sorting the layer by accuracy, and the result is $S = \{2, 4, 5, 6, 7\}$ and $W = \{6\}$. Fig. 2 (b) illustrates the depth prediction accuracy with respect to the number of training images and evaluated timesteps. We chose 100 images for the depth predictor since the accuracy gain from using more images is relatively small. We also show the scalability of our depth predictor with the latent diffusion [19] backbone in Fig 3.

Quantitative results. We compare the results of evaluation metrics for LSUN-Bedroom and LSUN-Church [31] between unguided ADM [7] and ADM with our guidance, and the results are shown in Tab. 1 and 2 respectively. The results indicate that ADM guided by DPG or DCG shows better performance than our baseline in dFID, a metric used to evaluate performance in the depth domain. In the appendix, we show that FID is inadequate for measuring depth-awareness.

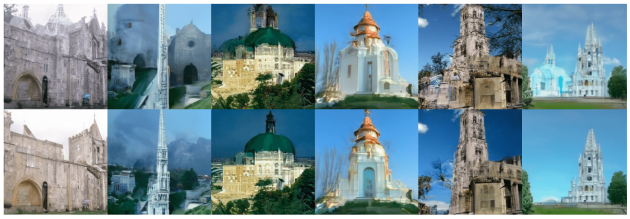


Figure 6. **Qualitative results on LSUN-church [31] dataset.** First row is unguided samples from DDIM [26], and the second row is guided samples using our guidance method, called depth-aware guidance (DAG).

Method	$\delta > 1.25$ (↑)	AbsRel. (↓)
Supervised	79.06	0.144
Unguided data	72.55	0.185
DAG-based data	**77.54**	**0.151**

Table 3. **Application of learning monocular depth estimation.** We train the depth estimation model from scratch using U-Net [20] based backbone network with our synthesized data.

Qualitative results. We compare the result from ADM with and without our guidance method in Fig. 4. We show both generated images and predicted depth maps using DPT, which demonstrates more robust depth prediction with our method. We also demonstrate the effectiveness of our method in 3D scene understanding through surface normal estimation and point cloud visualization (Fig. 5). Compared to the baseline, our predictions have clearer boundaries and higher level of detail. Fig. 6 shows qualitative comparisons on LSUN-Church, where our guidance method preserves geometric characteristics effectively.

Application for monocular depth estimation To improve the effects of our generation as unlabeled data, we leverage the guided images and corresponding depth maps. We train the U-Net-based depth estimation network [8] and evaluate the metrics with the NYU-Depth datasets [23]. We compare the training results using reference data, unguided generated results, and our generated results. For the depth evaluation, we use accuracy under the threshold ($\delta < 1.25$) and absolute relative error (AbsRel). Tab. 3 shows that the images generated by DAG-based data are more helpful in training the depth predictor than the unguided samples set.

4. Conclusion

In this paper, we propose a label-efficient method for predicting depth maps of images generated by the reverse process of diffusion model using internal representations. We also introduce a novel guidance scheme to guide the image to have a plausible depth map. In addition, we present a evaluation metric that effectively represents depth awareness using pretrained depth estimation networks.

References

[1] Emmanuel Brempong Asiedu, Simon Kornblith, Ting Chen, Niki Parmar, Matthias Minderer, and Mohammad Norouzi. Decoder denoising pretraining for semantic segmentation. *arXiv preprint arXiv: Arxiv-2205.11423*, 2022. 1

[2] Philip Bachman, Ouais Alsharif, and Doina Precup. Learning with pseudo-ensembles. *Advances in neural information processing systems*, 27, 2014. 1, 2

[3] Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13137–13146, October 2021. 3

[4] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 1, 2

[5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15750–15758, 2021. 3

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 2

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 1, 3, 4, 7

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3828–3838, 2019. 4

[9] Alexandros Graikos, Nikolay Malkin, Nebojsa Jojic, and Dimitris Samaras. Diffusion models as plug-and-play priors. *arXiv preprint arXiv:2206.09012*, 2022. 1, 3

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 1

[11] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 1

[12] Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. *arXiv preprint arXiv:1610.02242*, 2016. 1

[13] Dong-Hyun Lee et al. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Workshop on challenges in representation learning, ICML*, volume 3, page 896, 2013. 2

[14] 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 *International Conference on Learning Representations*, 2021. 3

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

[16] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 1

[17] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023. 1, 3

[18] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 3, 4, 7

[19] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 1, 2, 4

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[21] Mehdi Sajjadi, Mehran Javanmardi, and Tolga Tasdizen. Regularization with stochastic transformations and perturbations for deep semi-supervised learning. *Advances in neural information processing systems*, 29, 2016. 1, 2

[22] Zifan Shi, Yujun Shen, Jiapeng Zhu, Dit-Yan Yeung, and Qifeng Chen. 3d-aware indoor scene synthesis with depth priors. *arXiv preprint arXiv:2202.08553*, 2022. 4

[23] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 4, 7

[24] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 1

[25] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *NeurIPS*, 33:596–608, 2020. 2, 3

[26] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 4

[27] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *arXiv preprint arXiv: Arxiv-1512.00567*, 2015. 7

[28] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6243–6252, 2017. 1

[29] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust

features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine learning*, pages 1096–1103, 2008. 1

[30] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1

[31] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1, 3, 4

Appendix: Self-Predicted Depth-Aware Guidance with Diffusion Models

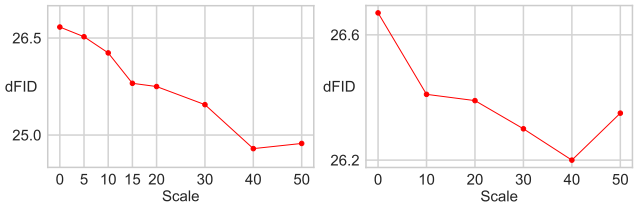
Guidance scale. We study the relationship between the guidance scale and image quality in terms of dFID for each DCG and DPG. Fig. 7 shows the tendency of the metrics versus scales of each guidance. As depicted in Fig. 7, DCG and DPG obtains the best results at $\omega_{\text{dc}} = 40$ and $\omega_{\text{dp}} = 40$ in dFID, respectively. Therefore, we treat this scale as default during the experiment.

Resolution of the prior network. In our second guidance method, DPG, we need to train a diffusion network to give a prior for condition image. We test three resolutions for the pretrained prior diffusion network, and the quantitative results are shown in Tab. 4. The 128×128 outperforms the other resolution in dFID. But due to limitations in computation cost for the sampling process, so we choose the 64×64 for the prior diffusion network. As our output depth map has a resolution of 64×64, we interpolate the depth map when fed to the prior network.

Depth FID. To compute the FID of the depth domain, we first estimate the depth map using the NYU-Depth [23] pretrained depth estimator DPT-Hybrid [18], which is available in the official repository. Then, identical to the original FID, we compute the Fréchet distance of embeddings collected from depth images using the Inception v3 model [27]. We show examples in Tab. 5 where FID is good but the geometric realism is limited. It demonstrates that the FID is inappropriate for measuring geometrical awareness. To address this problem, we propose dFID. DAG, our proposed guidance method, improves both the visual quality of geometric realism and the dFID metric. Additionally, we observed a trade-off between dFID and FID to some extent. This can be interpreted as a trade-off between structure awareness and texture quality, which is similar to diversity-fidelity trade-off shown in previous works of guidance as in ADM [7].

Method	dFID (↓)	FID (↓)
Baseline	26.77	**18.24**
32×32	25.14	19.86
64×64	25.96	18.85
128×128	**25.07**	21.69

Table 4. **Ablation study of the resolution of the prior network.**



(a) DCG (b) DPG

Figure 7. **Comparison of dFID with respect to the guidance scales of DCG and DPG.**

Samples				
DAG			✓	
dFID (↓)	15.71		**13.93**	
FID (↓)	**6.72**		7.59	

Table 5. **Comparison of generated samples from ADM.** The images in the left two columns are examples of samples from ADM without our guidance method, DAG, and the images in the right two columns are examples from ADM using our method.