# More Than Generation: Unifying Generation and Depth Estimation via Text-to-Image Diffusion Models

**Hongkai Lin**    **Dingkang Liang**    **Mingyang Du**    **Xin Zhou**    **Xiang Bai**✉

Huazhong University of Science and Technology

{hklin,dkliang,xbai}@hust.edu.cn
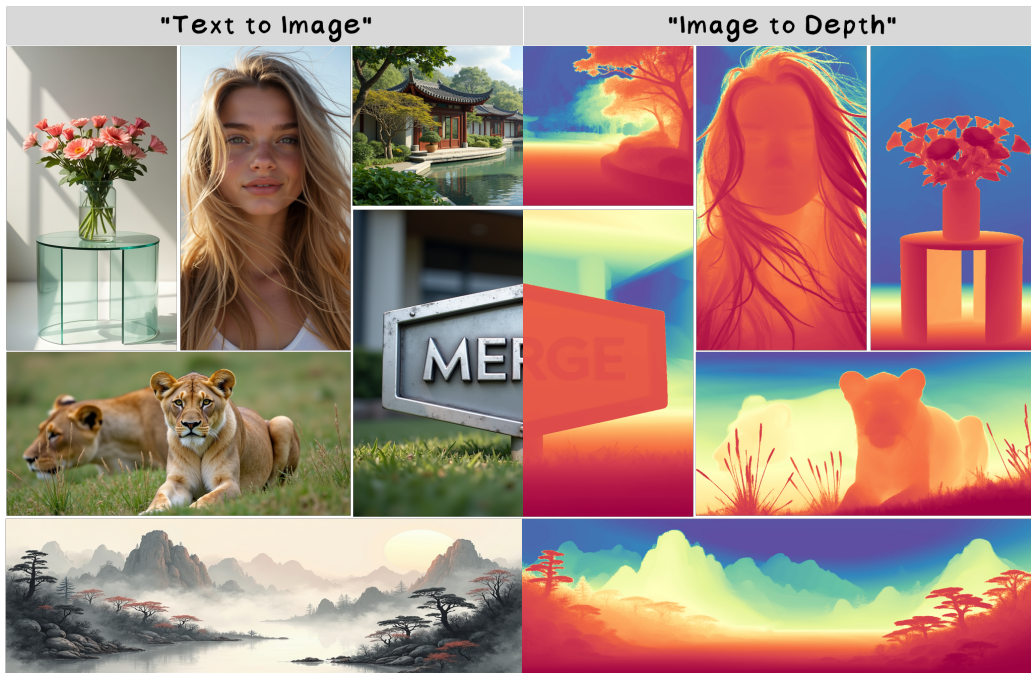
Project code: https://github.com/H-EmbodVis/MERGE

Figure 1: We present MERGE, a simple unified diffusion model for image generation and depth estimation. Its core lies in leveraging streamlined converters and rich visual prior stored in generative image models. Our model, derived from fixed image generation models and fine-tuned pluggable converters with synthetic data, expands powerful zero-shot depth estimation capability.

## Abstract

Generative depth estimation methods leverage the rich visual priors stored in pre-trained text-to-image diffusion models, demonstrating astonishing zero-shot capability. However, parameter updates during training lead to catastrophic degradation in the image generation capability of the pre-trained model. We introduce MERGE, a unified model for image generation and depth estimation, starting from a fixed pre-trained text-to-image model. MERGE demonstrates that the pre-trained text-to-image model can do more than image generation, but also expand to depth estimation effortlessly. Specifically, MERGE introduces a play-and-plug framework that enables seamless switching between image generation and depth estimation

✉ Corresponding author.

modes through simple and pluggable converters. Meanwhile, we propose a Group Reuse Mechanism to encourage parameter reuse and improve the utilization of the additional learnable parameters. MERGE unleashes the powerful depth estimation capability of the pre-trained text-to-image model while preserving its original image generation ability. Compared to other unified models for image generation and depth estimation, MERGE achieves state-of-the-art performance across multiple depth estimation benchmarks.
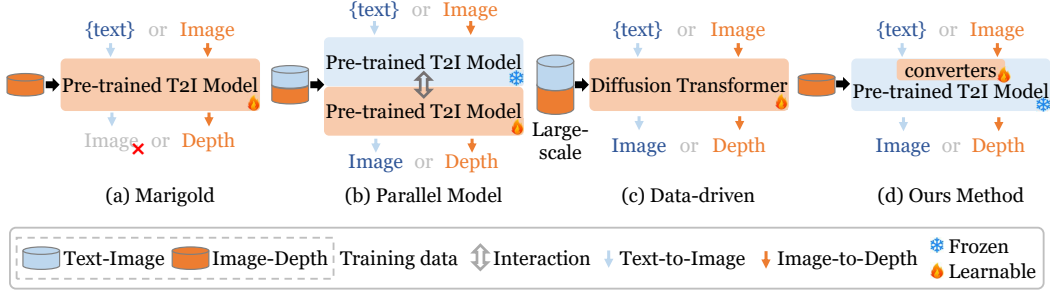


Figure 2: The comparison between existing methods and ours shows that, unlike previous works, our method requires only a few additional parameters to unleash its powerful depth estimation capability without compromising its inherent T2I generation ability.

# 1 Introduction

Diffusion models [12], especially text-to-image (T2I) models [1, 19, 38], have made astonishing progress in the quality of generated images. Meanwhile, emerging generative depth estimation research [10, 11, 16] reveals the latent potential for depth estimation tasks of advanced T2I models. Exploring a framework bridging image generation and depth estimation is an essential step toward building a Unified Model used for generation and visual perception.

Marigold [16], a pioneer of generative depth estimation, shows a powerful zero-shot depth estimation method leveraging pre-trained T2I diffusion models, as shown in Fig. 2(a). Even though the irreversible degradation of its original generation capability due to parameter updates during training, the paradigm still makes it possible to unify image generation and perception within diffusion models. In contrast, JointNet [52] and UniCon [21] present an alternative strategy by utilizing a parallel dual-model interaction architecture to unify generation and depth estimation tasks, as shown in Fig. 2(b). The latest work, OneDiffusion [20], introduces a data-driven solution, as shown in Fig. 2(c), to train a unified generation and perception model from scratch with massive multitask data (100M).

Despite remarkable progress, such approaches may be considered inefficient or resource-intensive solutions, potentially suboptimal choices. Considering rich visual prior is stored in advanced T2I models, a natural question arises: *Can T2I models effortlessly expand depth estimation capability without degrading their image generation capability?* This paper explores a unified diffusion model without bells and whistles for image generation and depth estimation tasks, starting from a fixed T2I model. Our method, referred to as **MERGE** (as shown in Fig. 2(d)), demonstrates that T2I models can do **M**ore than gen**ER**ate ima**GE**s. With a few simple converters, they can effortlessly unleash the powerful depth estimation capability.

To enable depth estimation ability in a fixed T2I diffusion model, we present a play-and-plug framework without bells and whistles. Before each transformer layer of the pre-trained T2I diffusion transformer (DiT), hereafter referred to as the T2I block, we introduce an identical, learnable T2I block as a converter. This converter transforms the latent features originally tailored for the T2I task into features suitable for the depth estimation task. This straightforward play-and-plug design enables seamless switching between the original image generation and depth estimation models by skipping or running these converters. Meanwhile, considering the similarity of output features between pre-trained T2I blocks, we propose a **G**roup **RE**use (GRE) mechanism to improve the utilization of the converter. Specifically, pre-trained T2I blocks are divided into several groups, and a shared within-group converter is used to effortlessly transform the T2I model into a generative

depth estimation model. During inference, if the text-to-image generation capability is needed, these additional converters can be easily skipped, restoring the depth estimation model to the T2I model. In addition, some empirical studies further simplify the converters with virtually no impact on MERGE's depth estimation performance. Combining these simple, flexible, and pluggable converters with the group reuse mechanism enables MERGE to unify image generation and depth estimation with no difficulty. Remarkably, MERGE presents a unified model starting from a fixed T2I model, requiring only a few additional parameters and fine-tuning on depth estimation datasets.

We conduct comprehensive evaluations of MERGE on established depth estimation benchmarks (NYUv2 [40], ScanNet [6], and DIODE [43]). The results show that MERGE achieves state-of-the-art performance across multiple metrics and benchmarks while introducing merely 12% additional trainable parameters. Especially on the NYUv2 benchmark, MERGE achieves 5.9 A.Rel and $95.4\%$ $\delta 1$, outperforming OneDiffusion, which is trained from scratch with massive data.

MERGE demonstrates a simple and effective approach to enabling pre-trained T2I diffusion models with depth estimation capability while preserving their inherent T2I generative ability. We hope our work provides valuable insights into the construction of unified models for generation and perception, while offering a cost-effective solution ($\sim 12\%$ additional parameters) for extending the capabilities of pre-trained image generation models. The main contributions of this work are as follows: **1)** We explore how to unleash the depth estimation capability from fixed pre-trained T2I models while preserving their original generation ability and propose a simple method without bells and whistles called MERGE. The key lies in a flexible play-and-plug framework that allows seamless switching between image generation and depth estimation modes. **2)** We present a streamlined converter that effortlessly transforms features suited for image generation into depth estimation. Considering the feature similarity between different layers in pre-trained T2I models, we propose a Group Reuse Mechanism, which encourages converter reuse to improve parameter efficiency.

## 2    Related Work

**Text-to-Image Diffusion Model.** Recently, text-to-image (T2I) generation models [1, 30, 39] based on the diffusion principle [12] make significant progress in image quality, diversity, and efficiency. Stable Diffusion [38], as the pioneering latent diffusion model, compresses the T2I generation process into the latent space using VAE [17], greatly improving computational efficiency and the quality of generated images. SDXL [32] introduces dual text encoders [15, 33] and fully upgrades Stable Diffusion, further enhancing the quality and resolution of generated images. VQ-Diffusion [9] proposes a vector quantized diffusion model by replacing VAE with VQ-VAE [41], addressing the unidirectional bias problem. Unlike these denoising diffusion models based on UNet architectures, PixArt [4, 5] introduces a novel text-to-image paradigm built upon Diffusion Transformers (DiT) [31]. By integrating vision-language models [27] to construct high-quality text-image pairs data, PixArt achieves performance comparable to advanced works at significantly reduced computational cost. MMDiT, an innovative Multimodal Diffusion Transformer architecture, is introduced by Stable Diffusion 3 [7] and demonstrates excellent image generation capability. The state-of-the-art method, FLUX [19], builds a larger T2I diffusion model by mixing MMDiT with the Single-DiT, presenting stunning image generation capability. Distinguishing these excellent T2I works, this paper focuses on exploring how to unleash the depth estimation capability of pre-trained T2I models (specifically those transformer-based architectures) while preserving their inherent T2I generation ability.

**Diffusion-based Depth Estimation Model.** Unlike discriminative perception methods [23, 24, 36, 45, 46, 48] and these works [22, 44, 47, 54] that treat T2I models as feature extractors, emerging research [14, 16, 25] reveals the compelling potential of pre-trained T2I models as generative depth estimation models. GeoWizard [8] leverages a geometry switcher inspired by Wonder3D [29] to extend a single stable diffusion model to produce depth and normal results. Meanwhile, a simple yet effective scene distribution decoupler strategy proposed in this work boosts the capture of 3D geometry. DepthFM [10] is the pioneer of combining Flow Matching [26] to explore generative monocular depth estimation, significantly improving inference speed while maintaining accuracy. BetterDepth [53] is a straightforward and powerful depth estimation framework that refines the results of generative depth estimation by using the output of a foundation depth estimation model [48, 49] as a prior. Lotus [11] introduces a tuning strategy called detail preserver that achieves more accurate and fine-grained predictions, especially on reflective objects. JointNet [52] and UniCon [21] propose a unified architecture for image generation and depth estimation, leveraging parallel dual diffusion

(a) Training phase of our MERGE
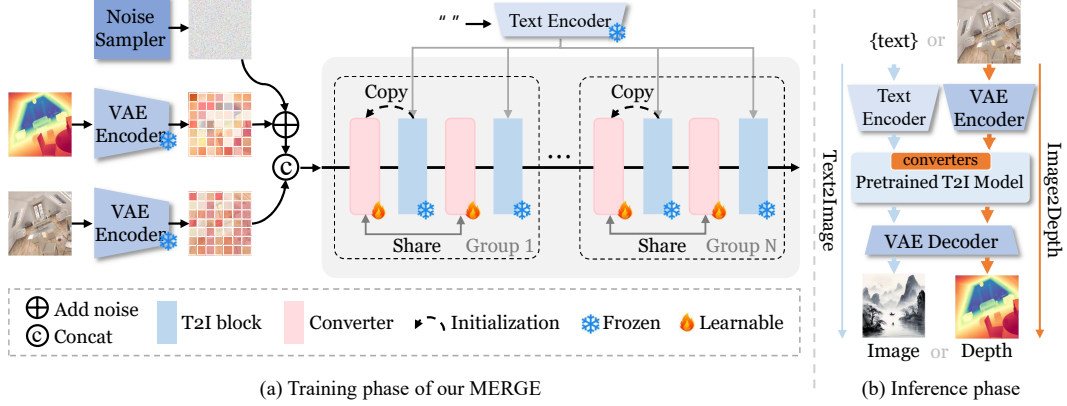(b) Inference phase

Figure 3: The pipeline of MERGE. Starting from the fixed DiT-based text-to-image (T2I) model, where transformer layers (hereafter referred to as T2I blocks) are divided into different groups. A shared and learnable converter is inserted before each T2I block within a group, transforming it into a depth estimation model. It can be reverted to the original T2I model by skipping these converters.

models with feature interaction operations between them. The latest work, OneDiffusion [20], presents a massive data-driven unified model for image generation and depth estimation. In contrast to existing full-parameter fine-tuning methods, which catastrophically degrade the image generation ability of pre-trained T2I models, or resource-intensive approaches, we explore a method that preserves the inherent generation capability of the T2I model while seamlessly transforming it into a generative depth estimation model.

## 3 Preliminaries

In this section, we review latent text-to-image diffusion models, with Stable Diffusion [38] serving as a representative of the latent diffusion paradigm.

Diffusion Models (DMs) [12] establish advanced text-to-image generation frameworks renowned for synthesizing photorealistic images through iterative denoising. DMs learn to reconstruct complex data distributions by approximating the reverse of a predefined diffusion process. Denoting $z_t$ as the random variable at the $t$-th timestep, the diffusion process is modeled as a Markov Chain:

$$z_t \sim \mathcal{N}(\sqrt{\alpha_t}z_{t-1}, (1 - \alpha_t)\boldsymbol{I}), \tag{1}$$

where $\alpha_t$ is a fixed coefficient predefined in the noise schedule, and $\boldsymbol{I}$ refers to the identity matrix. A prominent variant, the Latent Diffusion Model (LDM) [38], innovatively shifts the diffusion process of standard DMs into a latent space. This transition notably decreases computational costs while preserving the generative quality and flexibility of the original model. The resulting efficiency gain primarily arises from the reduced dimensionality of the latent space, which allows for lower training costs without compromising the model's generative capability.

Stable Diffusion, an exemplary implementation of LDM, comprises an AutoEncoder [41] and a latent denoising model. The AutoEncoder $\varepsilon$ is designed to learn a latent space that is perceptually equivalent to the image space. Meanwhile, the LDM $\epsilon_\theta$ is parameterized as a denoising model with the multimodal feature interaction module and trained on a large-scale dataset of text-image pairs via:

$$\mathcal{L}_{LDM} := \mathbb{E}_{\varepsilon(x),y,\epsilon \sim N(0,1),t}[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2], \tag{2}$$

where $\epsilon$ is the target noise, and $x$ is an RGB image. $\tau_\theta$ and $y$ are the pre-trained text encoder (e.g., CLIP [33], T5 [34]) and text prompts, respectively. This equation represents the mean-squared error (MSE) between the target noise $\epsilon$ and the noise predicted by the model, encapsulating the core learning mechanism of the latent diffusion model. Some of the latest text-to-image works (e.g., Stable Diffusion 3.5 [7], FLUX [19]) attempt to introduce more advanced flow matching [26] as an optimization objective, as it can more directly establish a mapping between the noise and data distributions, further improving image generation quality and efficiency.
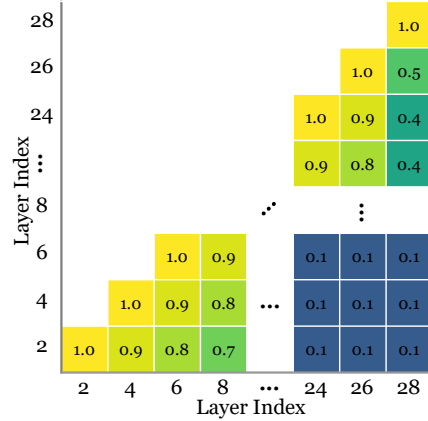
4

# 4 Our MERGE

To unleash the potential depth estimation capability of pre-trained text-to-image (T2I) models while retaining their inherent image generation capability. We present MERGE, as shown in Fig. 3(a), which requires only simple modifications to the pre-trained T2I model that can effortlessly unify image generation and depth estimation. The core of MERGE lies in a play-and-plug framework and the Group Reuse Mechanism without bells and whistles, designed for unified image generation and depth estimation. In addition, we present important empirical investigations, which significantly reduce the number of learnable parameters while keeping depth estimation capability virtually unaffected.

## 4.1 Play-and-Plug Framework

The T2I diffusion model learns the training data distribution to generate infinite data. Among these infinite pixel-level combinations that form images, there may exist underlying information highly correlated with the depth estimation modality. However, due to weak prompts and the sparse representation of depth information within the overall data distribution, extracting this latent feature of depth estimation modality from a fixed pre-trained T2I model is challenging.

In contrast to previous works [10, 11, 16] on fine-tuning the denoising models with full parameters, which catastrophically disrupt the inherent generation ability, we introduce play-and-plug, learnable converters before each T2I block in the pre-trained T2I model to guide and unleash its potential depth estimation capability. This play-and-plug converter design more naturally leverages the visual priors stored in the pre-trained T2I model, presenting highly comparable or even better depth estimation capability than full-parameter fine-tuning with fewer training parameters.



Features similarity between T2I blocks

Figure 4: The cosine similarity between the output features of different T2I blocks within the PixArt [5] model.

This straightforward strategy, without bells and whistles, demonstrates remarkable depth estimation capability and preserves the inherent image generation capability of the pre-trained T2I model, which is an aspect that previous works do not explore.

## 4.2 Group Reuse Mechanism

The solution of copying a learnable block before each pre-trained T2I block as a converter is simple yet effective. However, this method, which can be seen as directly coupling two models, does not efficiently leverage the powerful visual knowledge of the pre-trained T2I model to seamlessly guide its latent depth estimation capability, making it an inefficient approach. Interestingly, we observe that the closer two T2I blocks in the pre-trained T2I model are more similar in their output features in most cases, as shown in Fig. 4. Considering this observation, we propose a Group Reuse Mechanism (GRE) to encourage converter reuse and improve parameter utilization. Specifically, the T2I block in the pre-trained denoising model is divided into different groups, as shown in Fig. 3(a). The T2I block in a group shares a converter, which transforms the latent features suitable for the image generation task into the depth estimation task. Since GRE considers the feature similarity between different T2I blocks, it significantly reduces the additional learnable parameter number at a minor performance cost.

The collaboration between the play-and-plug framework and the Group Reuse Mechanism forms MERGE, which efficiently and flexibly unleashes the powerful zero-shot depth estimation potential of the pre-trained T2I model while retaining its inherent image generation capability.
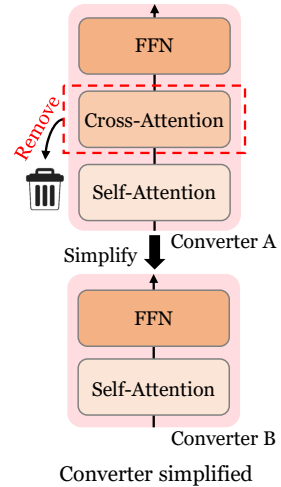


Converter simplified

Figure 5: The process of converter simplification, exemplified by PixArt [5].

## 4.3 Empirical Investigation

To smoothly transform the pre-trained T2I model into a generative depth estimation model through MERGE, the converter in MERGE is copied from the first T2I block of the respective group in the pre-trained T2I model, as illustrated by Converter A in Fig. 5. These converters are designed for T2I tasks and generally include components for multimodal feature interaction, such as Cross-Attention. However, existing diffusion-based generative depth estimation methods typically set the text prompt to empty. Intuitively, we suspect that the multimodal feature interaction design in the converter might be redundant for MERGE, as an empty text prompt contains no useful information. The experiments in Sec. 5.3 confirm this hypothesis, demonstrating that directly removing this multimodal interaction design, as illustrated by Converter B in Fig. 5, has almost no effect on the depth estimation performance of MERGE. This simplified converter reduces the 25% learnable parameter number in MERGE used to unleash depth estimation capability.

Similarly, in some of the latest MMDiT-based methods, like FLUX [19], a dual-stream multimodal interaction block can be simplified into a single-stream block to serve as a converter for MERGE. Using these simplified converters, MERGE can naturally leverage the rich visual priors of the pre-trained T2I model to demonstrate powerful zero-shot depth estimation capability while preserving its inherent image generation ability, which full-parameter fine-tuning methods cannot achieve.

## 5 Experiments

**Dataset and Evaluation metrics.** Following prior work [16, 20, 21], we train our model on the Hypersim [37] and Virtual KITTI [3] datasets, which correspond to synthetic indoor and outdoor datasets, respectively, comprising a total of 74k training samples. Subsequently, we evaluate its depth estimation performance on three real-world datasets: NYUv2 [40], ScanNet [6], and DIODE [43]. The quantitative evaluation metrics include absolute relative error (A.Rel) and the percentage of inlier pixels ($\delta1$) with thresholds of 1.25.

**Implementation Details.** We implement MERGE using PyTorch and employ PixArt-XL-2-512×512 and FLUX.1-dev as our pre-trained text-to-image (T2I) models. The training is solely conducted on depth estimation data to build a unified image generation and depth estimation model. Following previous work [16], we double the input channel of the patchify layer, enabling it to handle image conditions for generative depth estimation. During the depth estimation inference process, this patchify layer can seamlessly replace the patchify layer used in the T2I process. Regarding the group reuse mechanism, we adopt a default strategy of evenly dividing groups to avoid model-specific designs. Training our MERGE takes 30K iterations using a batch size of 32. We use the Adam optimizer with learning rates of 1e-4 and 3e-4 for PixArt and FLUX, respectively. Additionally, we follow the defaults provided in [16] for depth data preprocessing settings. For hyperparameters not mentioned in the MERGE training process, we follow the fine-tuning settings provided by the official pre-trained T2I model. All of our experiments are implemented on 8 NVIDIA H20 GPUs.

### 5.1 Main Results

We present two versions of MERGE with different parameter scales using the pre-trained PixArt [5] and FLUX [19] models, referred to as MERGE-B and MERGE-L, respectively. Unless otherwise specified, the number of groups for MERGE-B and MERGE-L is set to 14 and 10, respectively.

**Compared with the state-of-the-art.** As shown in Tab. 1, MERGE-B achieves superior performance over JointNet [52] and UniCon [21] with only 110M additional learnable parameters. Distinguishing these methods that require running dual diffusion models in parallel, the play-and-plug framework of MERGE is more computationally efficient, as it only injects fewer converters in a sequential manner when needed. Notably, due to the feature interaction design between the parallel diffusion models, it registers new knowledge into the pre-trained T2I model, which affects its original image generation capability, whereas MERGE does not.

Compared to the latest work, OneDiffusion [20], which is driven by 100M data, our MERGE-L achieves superior performance, surpassing OneDiffusion by 0.9 A.Rel on NYUv2 and $1.1\%$ $\delta1$ on DIODE, while requiring only approximately 12% additional parameters of the pre-trained T2I model. Remarkably, the learnable parameter of MERGE needs only about half the learnable parameters of OneDiffusion. By leveraging the powerful visual knowledge stored in the fixed pre-trained T2I

Table 1: Quantitative comparisons on zero-shot depth estimation. We compare our MERGE with works capable of both image generation and depth estimation on zero-shot depth estimation benchmarks. "-B" and "-L" refer to MERGE based on the pre-trained PixArt [5] and FLUX [19] models, respectively. "#Param." refers to the learnable parameter number, and "($\cdot$%)" in "#Param." represents the proportion of trainable parameters relative to the size of the original model. **Bold** numbers are the best.

| Method | Reference | Training Data | #Param. | NYUv2 | | ScanNet | | DIODE | |
|---|---|---|---|---|---|---|---|---|---|
| | | | | A.Rel $\downarrow$ | $\delta1 \uparrow$ | A.Rel $\downarrow$ | $\delta1 \uparrow$ | A.Rel $\downarrow$ | $\delta1 \uparrow$ |
| *These discriminative methods have* **depth estimation** *capability only.* | | | | | | | | | |
| DPT [35] | ICCV 21 | 1.2M | 123M(100%) | 9.8 | 90.3 | 8.2 | 93.4 | 18.2 | 75.8 |
| HDN [51] | NeurIPS 22 | 300K | 123M(100%) | 6.9 | 94.8 | 8.0 | 93.9 | 24.6 | 78.0 |
| DepthAnything [48] | CVPR 24 | 63.5M | 335M(100%) | 4.3 | 98.1 | 4.2 | 98.0 | 27.7 | 75.9 |
| DepthAnythingv2 [49] | NeurIPS 24 | 62.5M | 1.3B(100%) | 4.4 | 97.9 | - | - | 6.5 | 95.4 |
| *These generative methods have* **depth estimation** *capability only.* | | | | | | | | | |
| Marigold [16] | CVPR 24 | 74K | 889M(100%) | 5.5 | 96.4 | 6.4 | 95.1 | 30.8 | 77.3 |
| GeoWizard [8] | ECCV 24 | 280K | 889M(100%) | 5.2 | 96.6 | 6.1 | 95.3 | 29.7 | 79.2 |
| DepthFM [10] | AAAI 25 | 63K | 889M(100%) | 6.5 | 95.6 | - | - | 22.5 | 80.0 |
| Lotus [11] | ICLR 25 | 59K | 889M(100%) | 5.4 | 96.8 | 5.9 | 95.7 | 22.9 | 72.9 |
| *These generative methods support both* **image generation** *and* **depth estimation**. | | | | | | | | | |
| JointNet [52] | ICLR 24 | 65M | 889M(100%) | 13.7 | 81.9 | 14.7 | 79.5 | - | - |
| UniCon [21] | ICLR 25 | 16K | 125M(15%) | 7.9 | 93.9 | 9.2 | 91.9 | - | - |
| OneDiffusion [20] | CVPR 25 | 100M | 2.8B(100%) | 6.8 | 95.2 | - | - | **29.4** | 75.2 |
| MERGE-B (ours) | - | 74K | 110M(18%) | 7.5 | 94.2 | 9.9 | 89.8 | 32.5 | 74.9 |
| MERGE-L (ours) | - | 74K | 1.4B(12%) | **5.9** | **95.4** | **7.1** | **94.0** | 31.4 | **76.3** |

model, MERGE unifies image generation and depth estimation with less than one-thousandth of OneDiffusion's training data scale. Moreover, MERGE demonstrates overall superior depth estimation capability compared to OneDiffusion.

In addition, in the qualitative comparison, as shown in Fig. 6, MERGE demonstrates superior visual results, particularly in hollow and reflective regions, as highlighted by the black boxed areas.

**Compared with efficient low-rank fine-tuning methods.** We also present the comparison results between the efficient low-rank fine-tuning method and MERGE based on the same pre-trained T2I model. To ensure a fair comparison, we adjust the rank to match the number of learnable parameters in MERGE-B. The results shown in Tab. 2, compared to LoRA [13] and DoRA [28], MERGE-B performs better on all datasets. Unlike LoRA and DoRA, which fine-tune parameters at a finer-grained layer level, MERGE shows a structured converter at a higher level to enable depth estimation ability and proposes a group reuse mechanism to improve parameter utilization.

**Compared with full-parameter fine-tuning.** To further demonstrate the excellent zero-shot depth estimation capability of MERGE, Tab. 3 presents a comparison with the Marigold [16] paradigm, a generative depth estimation approach that full-parameter fine-tunes the pre-trained T2I diffusion model. We use the same pre-trained T2I model to present the PixArt version of Marigold (Marigold-P). As shown in Tab. 3, MERGE-B achieves results highly comparable to the full-parameter fine-tuning used by Marigold while requiring only 18% additional learnable parameters. Notably, if the converter is injected before each pre-trained T2I block, referred to as MERGE-B-28, it performs better on the NYUv2 dataset with only 37% learnable parameters of Marigold-P. More importantly, MERGE retains the image generation capability of the pre-trained T2I model thanks to the play-and-plug framework, which the Marigold paradigm can not achieve due to full-parameter fine-tuning disrupting the original feature distribution.

**Compared with traditional discriminative methods.** In terms of quantitative performance, state-of-the-art traditional discriminative methods (e.g., DepthAnything v2 [49]) still lead in monocular depth estimation, as shown in Tab. 1, particularly on challenging benchmarks like DIODE [43] that

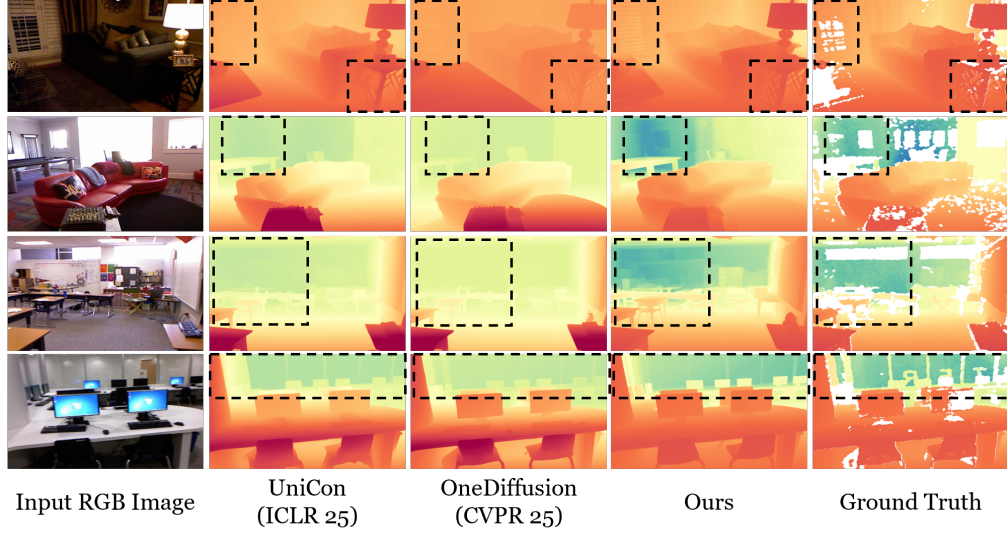|  | Input RGB Image | UniCon (ICLR 25) | OneDiffusion (CVPR 25) | Ours | Ground Truth |

Figure 6: Qualitative comparison between MERGE and other unified image generation and depth estimation methods on the NYUv2 benchmark. MERGE shows better depth estimation results, especially in detailed areas (such as free-form voids and reflect regions).

Table 2: Compared with low-rank fine-tuning methods based on the same text-to-image model [5].

| Method | Rank | #Param. | NYUv2 | | ScanNet | | DIODE | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | | | A.Rel ↓ | $\delta 1 \uparrow$ | A.Rel ↓ | $\delta 1 \uparrow$ | A.Rel ↓ | $\delta 1 \uparrow$ |
| LoRA [13] | 128 | 110M | 8.7 | 92.3 | 10.8 | 88.0 | 32.9 | 73.9 |
| DoRA [28] | 128 | 110M | 8.6 | 92.4 | 10.6 | 88.4 | 32.8 | 74.3 |
| MERGE-B | - | 110M | **7.5** | **94.2** | **9.9** | **89.8** | **32.5** | **74.9** |

Table 3: Compared with full-parameter fine-tuning. For a fair comparison, we adopt the same pre-trained T2I model to present the PixArt variant of Marigold [16], referred to as Marigold-P. "-28" indicates the number of groups is 28.

| Method | Support Tasks | | #Param. | NYUv2 | | ScanNet | | DIODE | |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| | Generation | Depth | | A.Rel ↓ | $\delta 1 \uparrow$ | A.Rel ↓ | $\delta 1 \uparrow$ | A.Rel ↓ | $\delta 1 \uparrow$ |
| Marigold-P | ✗ | ✓ | 596M | 7.4 | 94.2 | 9.5 | 90.1 | **31.5** | **76.0** |
| MERGE-B | ✓ | ✓ | 110M | 7.5 | 94.2 | 9.9 | 89.8 | 32.5 | 74.9 |
| MERGE-B-28 | ✓ | ✓ | 224M | **7.0** | **94.7** | **9.2** | **91.1** | 31.8 | 75.5 |

include diverse outdoor scenes. However, it is important to note that the performance gap narrows significantly on indoor datasets, for example, our MERGE achieves highly competitive performance to DepthAnything v2 on the NYUv2 [40] benchmark (A.Rel: 4.4 *vs.* 5.9, $\delta 1$: 97.9% *vs.* 95.4%).

## 5.2 Generalization to Normal Estimation Task

To further present the generality of our MERGE, we extend it to the zero-shot surface normal estimation. Following prior work [11], we employ the Hypersim [37] and Virtual KITTI [3] datasets as training data. Specifically, we filter out samples in Hypersim with invalid annotations and combine them with 20K Virtual KITTI training data, resulting in 59K training samples. All training hyperparameters are consistent with those used for the depth estimation task. We quantitatively evaluate the normal estimation performance of MERGE on NYUv2 [40], ScanNet [6], iBims-1 [18], and Sintel [2], reporting mean angular error ($m.$) as well as the percentage of pixels with an angular error below $11.25°$.

8

Table 4: Quantitative comparisons on normal estimation. "-B" and "-L" refer to MERGE based on the pre-trained PixArt [5] and FLUX [19] models, respectively. **Bold** numbers are the best.

| Method | Reference | Support Tasks | | NYUv2 | | ScanNet | | iBims-1 | | Sintel | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Gen. | Normal | $m.\downarrow$ | $11.25°\uparrow$ | $m.\downarrow$ | $11.25°\uparrow$ | $m.\downarrow$ | $11.25°\uparrow$ | $m.\downarrow$ | $11.25°\uparrow$ |
| Marigold [16] | CVPR 24 | ✗ | ✓ | 20.9 | 50.5 | 21.3 | 45.6 | 18.5 | 64.7 | - | - |
| GeoWizard [8] | ECCV 24 | ✗ | ✓ | 18.9 | 50.7 | 17.4 | 53.8 | 19.3 | 63.0 | 40.3 | 12.3 |
| StableNormal [50] | SIGGRAPH 24 | ✗ | ✓ | 18.6 | 53.5 | 17.1 | 57.4 | 18.2 | 65.0 | 36.7 | 14.1 |
| Lotus [11] | ICLR 25 | ✗ | ✓ | **16.5** | **59.4** | **15.1** | **63.9** | **17.2** | **66.2** | **33.6** | 21.0 |
| MERGE-B (ours) | - | ✓ | ✓ | 21.6 | 52.1 | 21.8 | 50.1 | 19.7 | 63.5 | 41.4 | 15.2 |
| MERGE-L (ours) | - | ✓ | ✓ | 20.1 | 54.3 | 18.3 | 60.2 | 18.2 | **66.2** | 36.9 | **21.3** |

Table 5: Ablation of the composition of converters. Here, GRE is disabled (i.e., a unique converter is inserted before each T2I block). SA/CA means the Self-Attention/Cross-Attention.

| Setting | SA | CA | FFN | FFN Scale | #Param. | A.Rel $\downarrow$ | $\delta1\uparrow$ |
|---|---|---|---|---|---|---|---|
| A | ✓ | ✓ | ✓ | ×4 | 596M | 6.9 | 94.8 |
| B | ✓ | ✗ | ✓ | ×4 | 447M | 6.9 | 94.6 |
| C | ✗ | ✗ | ✓ | ×4 | 298M | 7.6 | 93.6 |
| D | ✓ | ✗ | ✗ | - | 149M | 7.4 | 94.2 |
| E | ✓ | ✗ | ✓ | ×1 | 224M | **7.0** | **94.7** |

As shown in Tab. 4, even compared with existing task-specific generative methods based on full-parameter fine-tuning, our MERGE-L achieves highly comparable or superior results across multiple benchmarks. More importantly, MERGE preserves the original image generation capability of the pre-trained text-to-image model, which others can not achieve due to full-parameter fine-tuning disrupting the original feature distribution.

## 5.3 Ablation Studies

Unless otherwise specified, we conduct ablation studies on PixArt [5] using the NYUv2 dataset [40], and the training hyperparameters remain consistent with those mentioned earlier.

**Ablation on the composition of the converter.** We first validate the hypothesis in Sec. 4.3 under the setting of not enabling GRE, where the multimodal interaction design, Cross-Attention, is redundant for MERGE's converters, as it fails to extract effective information from an empty text prompt. The results, shown in setting A and setting B of Tab. 5, demonstrate almost no performance difference between with and without Cross-Attention, while the latter reduces approximately $25\%$ the learnable parameter number. Meanwhile, we further explore the impact of other components in the converter, such as Self-Attention and the Feedforward Network (FFN). As demonstrated by the results of setting C and setting D, removing Self-Attention or FFN reduces the learnable parameters of the converter, and it significantly degrades the depth estimation performance.

It may be ascribed to these operations' sole focus on latent image features, and removing these components would significantly weaken the model's representational capacity. Additionally, an interesting finding is that reducing the expansion rate of the FFN from 4 to 1 in converters, as illustrated in setting E, causes almost no impact on performance. This significantly reduces the learnable parameter number of the converter by approximately $36\%$.

**Ablation on the Group Reuse Mechanism (GRE).** To demonstrate the effectiveness of the Group Reuse Mechanism, we divide two adjacent T2I blocks into a group, resulting in a total of 14 groups with no overlap between them. When the converter is fixedly inserted before the first T2I block of each group, i.e., without GRE, it causes a significant decrease in depth estimation performance (A.Rel from $7.0 \rightarrow 15.6$, $\delta1$ from $94.7 \rightarrow 78.8$), as shown in the first two rows of Tab. 6. However, when sharing a converter within the group, the same parameter number yields a significantly improved result (A.Rel from $15.6 \rightarrow 7.5$, $\delta1$ from $78.8 \rightarrow 94.2$).

Compared to inserting a converter before each pre-trained T2I block, GRE reduces the learnable parameter number by about half, with only a slight performance cost. Tab. 6 also presents the impact of different group divisions on depth estimation performance. The results show that as the number of divided groups increases, the depth estimation performance of MERGE gradually improves. It is reasonable since features between consecutive layers share similarities but still exhibit differences. With more groups, these differences are more effectively balanced. Considering the trade-off between computational cost and performance, we ultimately adopt the setting of 14 groups as the default configuration for MERGE-B.

Table 6: Ablation of Group Reuse Mechanism (GRE).

| Groups | GRE | #Param. | A.Rel ↓ | $\delta 1$ ↑ |
|--------|-----|---------|---------|---------|
| 28 | ✗ | 224M | **7.0** | **94.7** |
| 14 | ✗ | 110M | 15.6 | 78.8 |
| 14 | ✓ | 110M | 7.5 | 94.2 |
| 7 | ✓ | 56M | 7.8 | 93.5 |
| 4 | ✓ | 32M | 9.3 | 91.0 |

**Ablation on the number of converters.** We also demonstrate the impact of stacking varying numbers of converters on depth estimation performance. The results in Tab. 7 show that stacking multiple converters to create a larger one is ineffective and even has negative effects. We suspect this is due to the significant increase in model depth, making optimization extremely challenging.

Table 7: Ablation of converter number.

| Number | #Param. | A.Rel ↓ | $\delta 1$ ↑ |
|--------|---------|---------|---------|
| 1 | 110M | **7.5** | **94.2** |
| 2 | 224M | 7.6 | 93.8 |
| 3 | 335M | 8.2 | 92.9 |

**Ablation on the initialization methods of the converter.** We further investigate the impact of different initialization methods for the converter on the depth estimation performance. The results in Tab. 8 show that initializing the converter with the pre-trained T2I block (the first T2I block of each group) consistently outperforms random initialization. This initialization strategy enables the converter to process the input features seamlessly.

Table 8: Ablation of initialization type.

| Init. Type | A.Rel ↓ | $\delta 1$ ↑ |
|------------|---------|---------|
| Random | 7.9 | 93.5 |
| Pre-trained | **7.5** | **94.2** |

**Ablation on different text prompts.** We finally investigate the impact of different types of text prompts on the depth estimation performance of MERGE. Specifically, we conduct experiments using three types of prompts: the default empty prompt, a fixed "depth map" prompt, and dense captions generated by LLaVA-Lightning-MPT-7B [27]. As shown in Tab. 9, our results demonstrate that more specific captions provide a performance boost for our method. Considering the trade-off between the additional cost of obtaining dense captions and the resulting benefits, we utilize an empty prompt by default.

Table 9: Ablation of the text prompt.

| Prompt Type | A.Rel ↓ | $\delta 1$ ↑ |
|-------------|---------|---------|
| Empty prompt | 7.5 | 94.2 |
| "Depth map" | 7.4 | 94.3 |
| Dense caption | **7.3** | **94.4** |

## 6    Conclusion

This paper presents MERGE, a method that starts from a fixed pre-trained text-to-image (T2I) model and unleashes its depth estimation capability while preserving its inherent image generation ability. Specifically, MERGE introduces a unified model without bells and whistles, which can seamlessly switch between original image generation and depth estimation modes through pluggable converters. Meanwhile, through empirical research, MERGE presents a simple and effective converter. Moreover, considering the similarity of output features across pre-trained T2I blocks, MERGE proposes a Group Reuse Mechanism to encourage parameter reuse, significantly reducing the additional parameter number. Extensive experiments demonstrate the effectiveness of MERGE. MERGE presents a new perspective for unifying multiple generative tasks besides the existing approaches that rely on massive data. Meanwhile, this play-and-plug strategy offers a cost-effective solution for extending the functionality of existing generative models.

**Limitation.** Despite the promising results achieved, our method still has limitations. For example, performing semantic segmentation using MERGE is highly challenging. One possible solution is to map semantic IDs to the RGB space [20]. However, the randomness of the denoising model makes it difficult to achieve a stable and accurate segmentation result by one-shot ID mapping from generated RGB results. In addition, integrating bit encoding methods, e.g., LDMSeg [42], is incompatible with the VAE of existing pre-trained text-to-image diffusion models. We consider this an important direction for future exploration.

## References

[1] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024.

[2] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Proc. of European Conference on Computer Vision*, 2012.

[3] Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

[4] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-$\sigma$: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *Proc. of European Conference on Computer Vision*, 2024.

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *Proc. of Intl. Conf. on Learning Representations*, 2024.

[6] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2017.

[7] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Proc. of Intl. Conf. on Machine Learning*, 2024.

[8] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *Proc. of European Conference on Computer Vision*, 2024.

[9] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2022.

[10] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. Depthfm: Fast monocular depth estimation with flow matching. In *Proc. of the AAAI Conf. on Artificial Intelligence*, 2025.

[11] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *Proc. of Intl. Conf. on Learning Representations*, 2025.

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proc. of Advances in Neural Information Processing Systems*, 2020.

[13] 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. In *Proc. of Intl. Conf. on Learning Representations*, 2022.

[14] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

[15] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[16] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2024.

[17] Diederik P Kingma. Auto-encoding variational bayes. In *Proc. of Intl. Conf. on Learning Representations*, 2014.

[18] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proc. of European Conference on Computer Vision*, 2018.

[19] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024.

[20] Duong H. Le, Tuan Pham, Sangho Lee, Christopher Clark, Aniruddha Kembhavi, Stephan Mandt, Ranjay Krishna, and Jiasen Lu. One diffusion to generate them all. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

[21] Xirui Li, Charles Herrmann, Kelvin CK Chan, Yinxiao Li, Deqing Sun, Chao Ma, and Ming-Hsuan Yang. A simple approach to unifying diffusion-based conditional generation. In *Proc. of Intl. Conf. on Learning Representations*, 2025.

[22] Ziyi Li, Qinye Zhou, Xiaoyun Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Open-vocabulary object segmentation with diffusion models. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023.

[23] Dingkang Liang, Tianrui Feng, Xin Zhou, Yumeng Zhang, Zhikang Zou, and Xiang Bai. Parameter-efficient fine-tuning in spectral domain for point cloud learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[24] Dingkang Liang, Wei Hua, Chunsheng Shi, Zhikang Zou, Xiaoqing Ye, and Xiang Bai. Sood++: Leveraging unlabeled data to boost oriented object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[25] Hongkai Lin, Dingkang Liang, Zhenghao Qi, and Xiang Bai. A unified image-dense annotation generation model for underwater scenes. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2025.

[26] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *Proc. of Intl. Conf. on Learning Representations*, 2023.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. of Advances in Neural Information Processing Systems*, 2024.

[28] Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. Dora: Weight-decomposed low-rank adaptation. In *Proc. of Intl. Conf. on Machine Learning*, 2024.

[29] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, et al. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2024.

[30] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proc. of Intl. Conf. on Machine Learning*, 2022.

[31] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023.

[32] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *Proc. of Intl. Conf. on Learning Representations*, 2024.

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proc. of Intl. Conf. on Learning Representations*, 2021.

[34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 2020.

[35] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2021.

[36] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[37] 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 *Proc. of IEEE Intl. Conf. on Computer Vision*, 2021.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2022.

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Proc. of Advances in Neural Information Processing Systems*, 2022.

[40] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proc. of European Conference on Computer Vision*, 2012.

[41] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *Proc. of Advances in Neural Information Processing Systems*, 2017.

[42] Wouter Van Gansbeke and Bert De Brabandere. A simple latent diffusion approach for panoptic segmentation and mask inpainting. In *Proc. of European Conference on Computer Vision*, 2024.

[43] Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z Dai, Andrea F Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R Walter, et al. Diode: A dense indoor and outdoor depth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

[44] Weijia Wu, Yuzhong Zhao, Hao Chen, Yuchao Gu, Rui Zhao, Yefei He, Hong Zhou, Mike Zheng Shou, and Chunhua Shen. Datasetdm: Synthesizing data with perception annotations using diffusion models. In *Proc. of Advances in Neural Information Processing Systems*, 2023.

[45] Gangwei Xu, Xianqi Wang, Zhaoxing Zhang, Junda Cheng, Chunyuan Liao, and Xin Yang. Igev++: Iterative multi-range geometry encoding volumes for stereo matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[46] Gangwei Xu, Yun Wang, Junda Cheng, Jinhui Tang, and Xin Yang. Accurate and efficient stereo matching via attention concatenation volume. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[47] 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 *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2023.

[48] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition*, 2024.

[49] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Proc. of Advances in Neural Information Processing Systems*, 2024.

[50] Chongjie Ye, Lingteng Qiu, Xiaodong Gu, Qi Zuo, Yushuang Wu, Zilong Dong, Liefeng Bo, Yuliang Xiu, and Xiaoguang Han. Stablenormal: Reducing diffusion variance for stable and sharp normal. In *ACM Transactions ON Graphics*, 2024.

[51] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. In *Proc. of Advances in Neural Information Processing Systems*, 2022.

[52] Jingyang Zhang, Shiwei Li, Yuanxun Lu, Tian Fang, David McKinnon, Yanghai Tsin, Long Quan, and Yao Yao. Jointnet: Extending text-to-image diffusion for dense distribution modeling. In *Proc. of Intl. Conf. on Learning Representations*, 2024.

[53] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. Betterdepth: Plug-and-play diffusion refiner for zero-shot monocular depth estimation. In *Proc. of Advances in Neural Information Processing Systems*, 2024.

[54] Wenliang Zhao, Yongming Rao, Zuyan Liu, Benlin Liu, Jie Zhou, and Jiwen Lu. Unleashing text-to-image diffusion models for visual perception. In *Proc. of IEEE Intl. Conf. on Computer Vision*, 2023.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the section of conclusion.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: This paper does not involve theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide detailed information about experiments and the source code that can reproduce the reported results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release of code and data.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify all the training and test details in the main text.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We follow the convention in prior works and report the performance number on the standard benchmarks.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [Yes]

Justification: See experiments part.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: We follow the NeurIPS Code of Ethic.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: no societal impacts.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

   Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

   Answer: [NA]

   Justification: No such risks.

   Guidelines:

   - The answer NA means that the paper poses no such risks.
   - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
   - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
   - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

   Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

   Answer: [Yes]

   Justification: We properly credited the creators or original owners of assets (e.g., code, data, models), used in the paper and conformed the license and terms.

   Guidelines:

   - The answer NA means that the paper does not use existing assets.
   - The authors should cite the original paper that produced the code package or dataset.
   - The authors should state which version of the asset is used and, if possible, include a URL.
   - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
   - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
   - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
   - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We communicated the details of the code/model as part of their submission.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: [NA]

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.