# MUSE: Unlocking Timestep as Native Task Steering for One-Step Dense Prediction

Shuo Zhou[1], Zhaoxin Li[1], and Xiujuan Chai[2]*

[1] Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing, China. {zhoushuo,lizhaoxin}@caas.cn
[2] Key Laboratory of Agricultural Big Data, Ministry of Agriculture and Rural Affairs, Beijing, China. chaixiujuan@caas.cn

**Abstract.** Monocular dense prediction has recently seen remarkable success by repurposing pre-trained diffusion models. This opens a promising yet challenging avenue for more efficient multi-task learning paradigm. However, existing multi-task diffusion methods often introduce parameter-heavy adapters, experts, or learnable task tokens, leading to computational redundancy. In this paper, we reveal an inherent mechanism within one-step diffusion models: the native, fixed timestep positional embedding can be repurposed as an endogenous task steering signal. Based on this discovery, we propose Multi-task Unified eStimation via timestep Embedding (MUSE), a parameter-free, single-model multi-tasking approach for dense prediction. We interpret this mechanism via Manifold Decoupling, where discrete, fixed timestep values deterministically steer the generation process towards decoupled, task-specific manifolds in the latent space. Extensive experiments across 10 datasets demonstrate that MUSE achieves highly competitive performance on both monocular depth and normal estimation, and its efficacy generalizes across U-Net and DiT architectures. Our work offers a concise and efficient path toward generalist vision models by simply unlocking the latent potential of existing generation infrastructure.

**Keywords:** One-step Diffusion · Monocular Dense Prediction · Multi-task Learning · Manifold Learning

## 1 Introduction

Geometric estimation tasks form the bedrock of spatial intelligence for autonomous agents. A machine's ability to perceive and reason about the 3D world from a single RGB image is critical for its large-scale applications. In recent years, the field has witnessed a paradigm shift with the advent of text-to-image (T2I) diffusion models [1–4]. Pre-trained on web-scale datasets [5], these models have developed an encyclopedic prior of the natural world. Consequently, a

* Corresponding author.

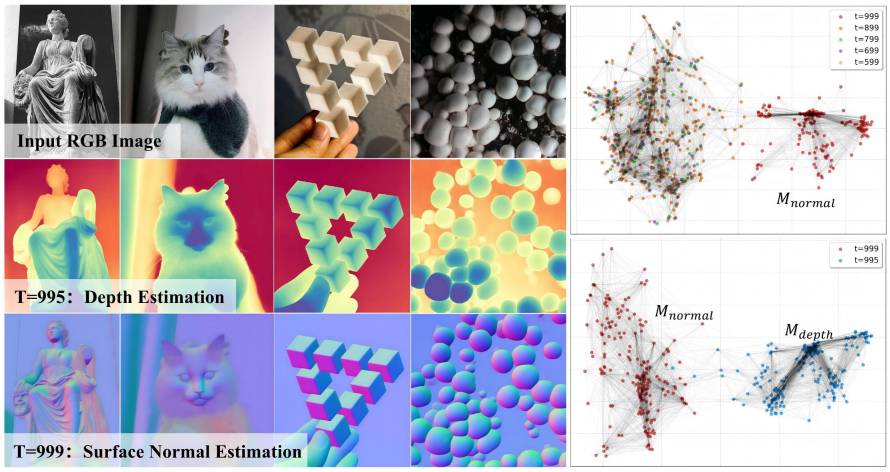


**Fig. 1:** We present MUSE, a single-model multi-tasking framework based on one-step diffusion for monocular dense prediction. Left: Qualitative overview of zero-shot depth and normal estimation by MUSE in challenging and ambiguous scenes. Right: Isomap 2D visualization of sample's latent features produced by a MUSE model which is steered with timestep $t_{\text{depth}}$=995 and $t_{\text{normal}}$=999. Features conditioned by steering timesteps are decoupled to distinct manifolds and exhibit clear separation, whereas those without being steered remain entangled, proving the Manifold Decoupling property, see Sec 4.3 for more details.

burgeoning branch of research has focused on *Repurposing* these powerful priors for visual perceiving tasks. Since the pioneering research of [6–8], taming conditional latent diffusion models (LDMs) [3] for monocular depth and normal estimation has been extensively explored [9–13]. The rich knowledge learned by the pre-trained VAE enables them to achieve competitive zero-shot performance by fine-tuning on a small number of high-quality synthetic data, demonstrating remarkable data efficiency and generalization capabilities.

Naturally, the success in single-task scenarios motivates the next frontier: developing a unified model for diverse dense prediction. However, extending diffusion to a multi-task setting is non-trivial. Prevailing approaches often resort to explicit architectural modifications, such as multi-head decoders [14], parameter-heavy Mixture-of-Experts (MoE) [15], or the injection of additional learnable task tokens and adapters [16–18]. Furthermore, attempting to jointly optimize a single shared network for disparate tasks often leads to gradient conflicts and negative transfer [19–21]. This raises a fundamental question: Must we introduce new parameters and complex modules to build a unified perception model, or does the pre-trained diffusion infrastructure already possess the latent capacity to decouple multiple tasks?

In this paper, we shift the focus from designing external modules to mechanism discovery. We present a counter-intuitive but remarkably effective insight: within a one-step diffusion framework, the native, fixed sinusoidal timestep em-

bedding — originally designed solely to indicate noise levels — can be repurposed as an endogenous and parameter-free task steering signal. Based on this, we propose the Multi-task Unified eStimation via timestep Embedding (MUSE). MUSE operates under a single-model multi-tasking paradigm in a truly minimalist fashion: by simply assigning discrete, constant timestep values to different tasks during fine-tuning and inference, it conditions a single, fully parameter-shared U-Net or DiT [22] to activate corresponding task-specific capabilities. MUSE achieves this without introducing any extra learnable task tokens, specialized loss functions, or considerable structural changes compared to the vanilla diffusion-based dense prediction models. Moreover, to elucidate the underlying mechanics, we provide a geometric interpretation termed *Manifold Decoupling*. By assigning orthogonal keys to each task, their respective optimization trajectories are fundamentally decoupled in the latent geometric space, elegantly avoiding the gradient conflicts that cause negative transfer. In this paper,"orthogonal" denotes functional disjointness and subspace decoupling, not zero dot-products.

Our main contributions are summarized as follows:

- For the first time we reveal a novel endogenous mechanism in one-step diffusion paradigm: the native, fixed timestep can function as a parameter-free semantic switch to steer distinct prediction tasks.
- MUSE, an elegant fine-tuning and inference protocol for unified monocular depth and normal estimation is proposed. Extensive experiments on 10 diverse benchmarks demonstrate that it achieves highly competitive performance, with generality across both U-Net and DiT architectures.
- We theorize and validate this mechanism through Manifold Decoupling, providing extensive geometric visualizations of the latent space separation.

## 2 Related Work

### 2.1 LDMs for Monocular Dense Prediction

The remarkable success of pre-trained T2I diffusion models has inspired a new research paradigm: repurposing their powerful visual priors for monocular dense prediction [6–11, 16–18, 23–25].

**Multi-Step Diffusion.** Early milestone works [6, 9] formulated this task as a conditional image generation problem. They operated under a stochastic, multi-step paradigm, where a target prediction map is iteratively denoised from random Gaussian noise, conditioned on the input RGB image. While effective, these methods are computationally intensive and suffer from inherent randomness, often requiring test-time ensembling [26] to produce stable results.

**Single-Step Diffusion.** Subsequent research has identified that the deterministic nature of dense prediction tasks is misaligned with the stochastic, noise-to-data probabilistic generation process. DMP [23] reformulated the diffusion process as a discriminative mapping between the image and the prediction by eliminating the random noise input. Works like GenPercept [24], Lotus [10, 12], and DepthMaster [25] demonstrated that the multi-step iterative process is not

essential for better performance, and a one-step, noise-free prediction can achieve or even surpass the performance of multi-step methods, while being orders of magnitude faster.

Along with some most recent explorations of optimizing one-step generation models [27–30], a crucial question arises: if the iterative process is unnecessary, what role does the timestep play beyond its traditional function as a noise level indicator? By peeling back the mathematical abstraction and delving into the implementation details, we unlock the potential of the timestep embedding as a powerful, yet underexplored, conditioning signal.

### 2.2 LDMs for Multi-Task Learning

The application of diffusion models to multi-task learning (MTL) can be broadly viewed from two perspectives.

**Explicit MTL Architectures.** One line of work focused on designing explicit architectures to enable a single diffusion model to handle multiple tasks. Some adopted the commonly used One-backbone with Multi-head design to generate outputs for different tasks [14]. Others [15, 31] transformed the U-Net into a Mixture-of-Experts (MoE) architecture to select task-specific computational paths. Orchid [11] and TaskDiffusion [32] proposed joint learning schemes by unifying task labels into a shared representation space, where the former even retrained the VAE module in a joint color-geometry latent space. Diception [33] adapted a Diffusion Transformer via lightweight task embeddings and point prompt, achieving strong multi-task capability. A more detailed breakdown of the diffusion model's conditional mechanisms can be found in the Supp. Materials Sec.A.

**Implicit MTL Strategy.** Another insightful perspective rethought the standard training of a diffusion model as an implicit form of MTL, where each timestep represents a distinct denoising task [19,20,34]. This viewpoint disclosed a fundamental challenge: the optimization objectives for different timesteps, especially those with disparate noise levels, can conflict. This results in conflicting gradients and leads to negative transfer [35], a classic problem in MTL that impedes training convergence and degrades task performance. To mitigate this, ANT [19] proposed clustering timesteps and applying MTL optimization techniques at the cluster level, Min-SNR [20] introduced a weighting scheme to balance the contributions of different timesteps, while DeMe [34] decoupled the training by fine-tuning separate models for different timestep ranges and then merging them. The accessibility of these methods is limited by their reliance on massive computing power, complex and cumbersome models, or extremely large trainset.

### 2.3 LDMs and Manifold Learning

The manifold hypothesis [36] posits that high-dimensional data lies on a low-dimensional manifold, revealing how a model captures the intricate structure of data distributions. Several studies have leveraged this geometric perspective

to analyze diffusion models. Chung et al. [37] enforced manifold constraints by projecting the intermediate results back onto a manifold defined by physical constraints at each denoising step. Similarly, He et al. [38] preserved the manifold structure during guided diffusion by projecting the guidance signal onto the manifold's tangent space. Stanczuk et al. [39] compellingly showed that a trained diffusion model implicitly learns the intrinsic dimension of the data manifold, which can be estimated from the Jacobian of the score function. Jacobsen et al. [40] proposed injecting noise along the tangent space of the data manifold to better preserve its structure. Recently, the JiT [41] has brought manifold learning to the forefront of the research community's attention. These works provide the theoretical and mathematical foundation for our Manifold Decoupling interpretation.

## 3 Method

### 3.1 Preliminaries

**Classical Denoising Diffusion Probabilistic Models.** A standard DDPM [1] consists of a forward noising process and a reverse denoising process. The forward process gradually perturbs clean data $x_0$ over $T$ steps by adding Gaussian noise according to a predefined variance schedule $\{\beta_t\}_{t=1}^{T}$:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \tag{1}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. The reverse process aims to recover the clean data from pure Gaussian noise $\mathbf{x}_T \sim \mathcal{N}(0, \mathbf{I})$ by learning a neural network $\epsilon_\theta(\mathbf{x}_t, t)$ to predict the added noise $\epsilon$ at each timestep $t$ from the noisy input $\mathbf{x}_t$. The model is trained by optimizing a simplified objective:

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{t,\mathbf{x}_0,\epsilon}\left[||\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)||^2\right] \tag{2}$$

Inference is an iterative process that starts from $\mathbf{x}_T$ and denoises it step-by-step to obtain the final sample $\mathbf{x}_0$.

**Conditional Diffusion for Dense Prediction.** Modern approaches like Stable Diffusion (SD) [3] operate in a compressed latent space. It is provided by a pre-trained Variational Autoencoder (VAE) with a pair of encoder $\mathcal{E}$ and a decoder $\mathcal{D}$. For a dense prediction task, given an input RGB image $I$ and its corresponding ground-truth map $Y$, their latent representations are obtained as $\mathbf{z}^x = \mathcal{E}(I)$ and $\mathbf{z}^y = \mathcal{E}(Y)$, respectively. The task is then formulated as learning a conditional denoiser $\epsilon_\theta(z_t, t, c)$ that predicts the noise from the noisy target latent $z_t$, conditioned on $c$. For dense prediction, $c$ is the input image latent feature, $c = \mathbf{z}^x$.

**One-Step and Deterministic Paradigm.** Recent advancements, epitomized by Lotus [10], have demonstrated that for discriminative dense prediction tasks, the iterative multi-step denoising process is not essential. Instead, a more efficient single-step paradigm can be adopted, which simplifies the process in two

key ways. First, it opts for directly predicting the clean latent $\mathbf{z}^y$ (termed $x_0$-prediction) rather than the noise $\epsilon$ ($\epsilon$-prediction). Second, it collapses the training and inference into a single, fixed timestep, typically the maximum timestep $t = T$ ($T = 999$ in SD). In the deterministic variant, the random noise component is entirely removed from the input, and surprisingly achieves the best performance on both depth and nomral estimation tasks. The model, now denoted as $f_\theta$, learns a direct mapping from the conditional image latent $\mathbf{z}^x$ to the target prediction latent $\mathbf{z}^y$. The training objective simplifies to a direct regression:

$$\mathcal{L}_{\text{one-step}} = \mathbb{E}_{\mathbf{I},\mathbf{Y}} \left[ ||\mathbf{z}^y - f_\theta(\mathbf{z}^x, T)||^2 \right] \tag{3}$$

where $T$ is the fixed timestep embedding.

### 3.2 MUSE: Multi-task Unified Estimation via Timestep Embedding

The timestep embedding, once a dynamic indicator of noise levels, becomes a static and underutilized signal in one-step generation. This paper posits that this native and powerful conditioning channel can be repurposed as a semantic switch for multi-task learning.

**Repurposing Timestep as the Task Steering** We propose that a single SD U-Net can be taught to perform distinct tasks by associating each task with a unique, discrete timestep value. Instead of treating the timestep $t$ as a continuous variable representing noise levels, we re-contextualize it as a categorical task prompt. This allows us to decouple the learning processes of multiple tasks in a parameter-free and distillation-free manner.

Formally, we define a set of target tasks $\mathcal{S} = \{\text{Task}_1, ..., \text{Task}_K\}$, and pre-assign a unique timestep to each: $\tau_{\text{task}} \in \{t_1, t_2, ..., t_k\}$. In MUSE, we set the timestep values as $t_{\text{rgb}}$, $t_{\text{depth}}$, and $t_{\text{normal}}$ corresponding to the tasks of RGB reconstruction, affine-invariant depth estimation and normal estimation. The one-step prediction model $f_\theta$ is now conditioned on these task-specific timesteps:

$$\mathbf{z}^y_{\text{pred}} = f_\theta(\mathbf{z}^x, t_{\text{task}}) \tag{4}$$

This formulation elegantly transforms the single-task model into a multi-task learner without adding any new parameters or modules. The model learns to interpret the fixed embedding of these discrete timestep values as instructions to activate different computational pathways within its vast parameter space, each specialized for a particular task. The unified training objective for MUSE is a simple summation of the Mean Squared Error (MSE) loss for each task. For a given training sample ($I$, $Y_{\text{task}}$), where $task \in \mathcal{S}$, the loss is computed as:

$$\mathcal{L}_{\text{MUSE}} = \mathbb{E}_{(\mathbf{I},\mathbf{Y}_{\text{task}}),t_{\text{task}}} \left[ ||\mathcal{E}(\mathbf{Y}_{\text{task}}) - f_\theta(\mathcal{E}(\mathbf{I}), t_{\text{task}})||^2 \right] \tag{5}$$

MUSE requires no complex task-specific loss functions or sophisticated loss-balancing strategies. Basically, the decoupling is not achieved at the loss level, but at the input level, by providing orthogonal task steering ($t_{\text{k-1}} \neq t_{\text{k}}$) that guide the model to learn non-conflicting gradient paths for each task.

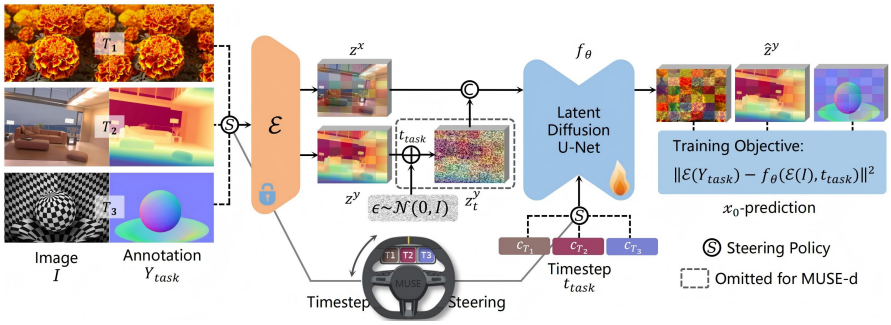


**Fig. 2:** Overview of MUSE finetuning protocol. In training stage, multiple fixed timesteps are sampled by a steering policy to condition the loss calculation. During inference, we can naturally use the timestep as a task steering to generate the corresponding prediction. By controlling the presence or absence of noise, generative or discriminative variant can be chosen.

**Training and Inference Protocol** The core of MUSE is the steering policy, involving a control of the data-loading and inference logic, as illustrated in Figure 2.

**Training.** Before all, the VAE of the pre-trained SD model is frozen, and only the U-Net parameters are fine-tuned. For each training iteration, we sample a mini-batch of training data as usual. While for every sample within the batch, the steering policy randomly endows it with one of the 3 task purposes: RGB reconstruction, depth estimation, or surface normal estimation. Then the task's pre-assigned timestep embedding is fed into the model along with the image latent feature. The model's output is then computed only against the corresponding ground-truth latent of the selected task. The total loss for the batch is the average of these individual sample MSE losses, and the gradients are backpropagated to update the single, shared U-Net parameters. In the baseline steering policy, the training frequency and loss contributions of sampled tasks is identical. It can also be adjusted by task ratio and loss weight based on the difficulty of the tasks and the characteristics of the datasets. This process forces the model to associate each specific timestep embedding with the goal of generating the corresponding task output.

**Inference.** The inference process fully showcases the efficiency and elegance of MUSE. To acquire predictions for all three tasks of a given input RGB image $I$, MUSE stacks the $I$ three times, assigns each $I$ a corresponding timestep, and then simultaneously outputs all the prediction maps in a single forward inference, thanks to the broadcast mechanism. If specific scenarios and computational resource limitations are taken into account, MUSE can also be used as a single-task model. For example, the model can be fed the $I$ along with timestep $t_{\text{depth}}$ for depth estimation task. The output latent is decoded to produce the depth map $\hat{Y}_{\text{depth}} = \mathcal{D}(f_\theta(\mathcal{E}(I), t_{\text{depth}}))$. The entire procedure is one-step for each task, requiring no iterative denoising. The ability to switch between tasks

by simply changing a single integer input makes MUSE an extremely efficient and flexible framework for unified visual perception. Some zero-shot depth and normal estimation are shown in Figure 1 and 4.

### 3.3 Formulation via Manifold Decoupling

The Manifold Hypothesis [39,42,43] states that high-dimensional real-world data, such as images of depth or surface normal, do not populate the entire ambient space but rather concentrate on or near a low-dimensional manifold embedded within it. Inspired by these principles, we contend that MUSE's success is rooted in the geometric structure of the data it learns to encode in latent space.

Let $\mathcal{Z}$ be the high-dimensional latent space of the VAE. The latent codes of all valid depth maps lie on a low-dimensional manifold $\mathcal{M}_{\text{depth}} \subset \mathcal{Z}$, and similarly, normal data lie on another distinct $\mathcal{M}_{\text{normal}} \subset \mathcal{Z}$. In a traditional multi-task learning scenario where parameters are shared through hard-coded methods, the model must learn to generate outputs on both manifolds simultaneously. This often leads to conflicting gradients: an update that moves a prediction closer to $\mathcal{M}_{\text{depth}}$ may inadvertently push it away from $\mathcal{M}_{\text{normal}}$, resulting in negative transfer.

The timestep embedding in MUSE acts as a geometric controllers that steers denoising dynamics along functionally segregated pathways. It conditions the function $f_\theta$ to learn a mapping that projects the image latent feature $\mathbf{z}^x$ specifically onto the manifold corresponding to that task. From this perspective, MUSE is not a single function, but a parameterized family of functions $\{f_{\theta,t}\}_{t\in\{t_{\text{task}}\}}$. Each $f_{\theta,t_{\text{task}}}$ is trained to approximate a projection onto the target task manifold:

$$f_\theta(\mathbf{z}^x, t_{\text{task}}) \approx \mathit{\Pi}_{\mathcal{M}_{\text{task}}}(\mathbf{z}^x) \tag{6}$$

Any update to a point $z_{\text{pred}} \in \mathcal{M}_{\text{task}}$ should ideally lie within the tangent space to ensure the updated point remains on the manifold [38]. Because the model learns to associate the orthogonal inputs $t_{\text{task}}$ with updates along their respective, and likely different manifolds, the optimization processes for different tasks do not interfere. MUSE learns distinct geometric objectives, guided by multiple distinct keys through the cross attention structure.

## 4 Experiments and Results

### 4.1 Implementation

We adopt the pre-trained weights of Stable Diffusion v2 [3] as the model initialization. All models are fine-tuned using the AdamW optimizer with a learning rate of $3 \times 10^{-5}$ and a total batch size of 16. The models are trained for 12K iterations on a single NVIDIA V100-32GB GPU. Following prevailing general pipelines [10, 26], we use a mixture of 9 : 1 high-quality synthetic datasets, Hypersim [44] and Virtual KITTI [45], for fine-tuning.

We mainly implement two variants. The generative paradigm MUSE-g follows the probabilistic diffusion setup, where a Gaussian noise is concatenated with the image latent $\mathbf{z}^x$ as input to the U-Net. The discriminative paradigm MUSE-d is deterministic and removes the explicit noise input, and the U-Net only receives the image latent $\mathbf{z}^x$ and the task timestep embedding $c_{\text{time}}$.

We validate our approach on 10 acknowledged dense prediction benchmarks. For affine-invariant depth estimation, we evaluate on NYUv2 [46], KITTI [47], ScanNet [48], ETD3D [49], and DIODE [50]. For surface normal estimation, we evaluate on the NYUv2, ScanNet, iBims-1 [51], Sintel [52], and Oasis [53] datasets. As widely recognized by the community, the reported performance metrics include: Absolute Relative Error (AbsRel↓) and $\delta_1 \uparrow$ accuracy for depth estimation, and Mean Angular Error (Mean↓) and the percentage of pixels with an error below 11.25° ↑ for surface normal estimation.

### 4.2 Mechanistic Analysis of Timestep-Driven Steering

To understand exactly how the scalar timestep governs the shared U-Net, we conduct controlled ablations to dissect the underlying mechanism. The design space of steering policy includes five dimensions: three timestep values corresponding to the three tasks, the frequency of task training, and the weights for loss calculation, as Table 1. We aim to answer two critical questions: Is the numerical distinctness of the timestep the sole driver of task decoupling, and does its absolute magnitude matter?

**The Necessity of Orthogonal Steering Policy.** As Figure 3 shows, we first establish the boundary conditions of negative transfer. When the model is forced to predict both depth and surface normals using an identical timestep ($t_{depth} = t_{normal}$), it suffers from severe gradient conflicts and catastrophic interference, failing to converge on either task. However, the exact same architecture achieves simultaneous convergence simply by assigning discrete, distinct values ($t_{depth} \neq t_{normal}$). This empirically validates our core premise: the model does not require extra parameters; it relies entirely on the orthogonality of the given timestep embeddings to steer input features to their corresponding, decoupled task manifolds.

**Table 1:** MUSE steer policy ablation study settings. In the last line, "tr" represents adjusting the training frequency (task ratio) of RGB reconstruction, depth estimation, and normal estimation tasks to 1:2.5:2.5. "lw" indicates that the weights of the three tasks (loss weight) are adjusted to 1:2.5:2.5 when calculating the loss.

| exp | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| g/d | generative | | | | | | | | | | | | | | | | discriminative | | | |
| $t_{rgb}$ | 599 | 599 | 599 | 599 | 599 | 599 | 599 | 599 | 599 | 799 | 799 | 991 | 997 | 997 | 997 | 997 | 997 | 991 | 997 | 997 |
| $t_{depth}$ | 599 | 799 | 999 | 699 | 799 | 899 | 999 | 999 | 999 | 899 | 999 | 995 | 998 | 998 | 998 | 998 | 998 | 995 | 599 | 999 |
| $t_{normal}$ | 599 | 799 | 999 | 999 | 999 | 999 | 699 | 799 | 899 | 999 | 899 | 999 | 999 | 999 | 999 | 999 | 999 | 999 | 799 | 999 |
| policy | - | - | - | - | - | - | - | - | - | - | - | - | - | tr | lw | trlw | tr | tr | tr | tr |

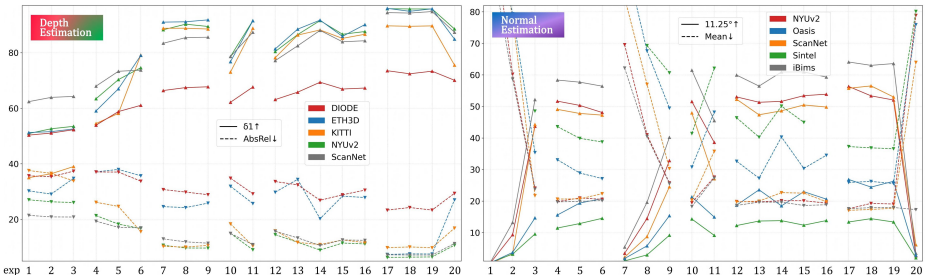


**Fig. 3:** MUSE steer policy ablation study evaluation metrics. The experiment (*exp*) numbers and groups of the horizontal axis correspond one-to-one with the 20 configurations in Table 1. The two metrics for each task are converted to [0,100] so they can be plotted in the same graph. Note that for every *exp*, there are 10 metrics for depth estimation in the left diagram, plus another 10 metrics for normal estimation in the right diagram. See Supp. Materials Sec.B for complete numerical data.

**From Noise Indicator to Categorical Semantic Switch.** We further investigate the network's sensitivity to specific numerical values of $t$. Interestingly, we observe a fundamental dichotomy governed by the presence or absence of explicit noise. In the probabilistic paradigm MUSE-g, where Gaussian noise is explicitly injected as the vanilla diffusion-based dense prediction models [9, 10], the model remains bound to the physical meaning of $t$ established during pre-training. Under the premise of following the classic DDPM method and huggingface implementation, we observe that higher timestep values (e.g., $t \to 1000$) yield superior performance. This aligns with the diffusion characteristic that high timesteps correspond to a latent space dominated by high-level semantic structures rather than low-level noise details, which is crucial for dense geometric reasoning. The best-performing MUSE-g variant utilizes the three highest timestep values ($t_{rgb} = 997$, $t_{depth} = 998$, $t_{normal} = 999$).

In stark contrast, in the deterministic paradigm MUSE-d, the explicit noise input is removed, and the physical interpretation of $t$ as a "noise-level indicator" is entirely voided. Our ablations reveal that MUSE-d is remarkably robust to the absolute magnitude of $t$. The timestep transitions purely into a discrete, categorical semantic switch. As long as the assigned keys are mutually distinguishable, their specific numerical magnitudes have minimal impact on task decoupling. This mechanistic transition confirms that in noise-free, one-step diffusion models, the native time embedding space is intrinsically rich enough to act as an endogenous multi-task router. In this paper, we mainly present MUSE-d with $t_{rgb} = 991$, $t_{depth} = 995$, $t_{normal} = 999$ alongside optimized data-sampling frequencies.

Based on the above findings and the optimal hyperparameter settings, we utilize a large batchsize 192 and spent 130 hours to train a MUSE-g and a MUSE-d model that aimed for the highest task performance under existing resources. The quantitative comparison with other methods are shown in Table 2. It is worth noting that MUSE achieves performance comparable to recent SOTA multi-task

**Table 2: Quantitative Comparison on Dense Prediction Methods.** We evaluate MUSE against SOTA methods across 10 diverse indoor and in-the-wild datasets. MUSE achieves highly competitive zero-shot generalization simultaneously on both depth and surface normal estimation. Note that for all methods that reported results with different inference steps, we focus only on the performance of one-step inference for fair comparison. "-" indicates the paper did not report results. The best-performing method for each metric is highlighted in **bold**. We use the same evaluation protocol as Lotus [10].

| Method | Gene- | MTL | NYUv2 | | ScanNet | | KITTI | | ETH3D | | DIODE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | rative | Model | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ |
| **Task 1: Affine-Invariant Depth Estimation** | | | | | | | | | | | | |
| Marigold | ✓ | × | 6.0 | 95.8 | 6.9 | 94.5 | 10.5 | 90.4 | 7.1 | 95.1 | 31.0 | 77.2 |
| Lotus-g | ✓ | × | 5.4 | 96.6 | 6.0 | 96.0 | 11.3 | 87.7 | 6.2 | 96.1 | - | - |
| Lotus-d | × | × | 5.3 | 96.7 | 6.0 | 96.3 | 9.3 | 92.8 | 6.8 | 95.3 | - | - |
| Marigoldv1.1 | ✓ | × | 5.9 | 96.1 | 6.6 | 95.3 | 11.0 | 88.8 | 7.0 | 95.5 | 30.4 | 77.3 |
| Lotus-2 | × | × | **4.1** | 97.6 | **4.2** | **97.6** | 6.7 | 94.5 | 4.6 | 97.6 | 22.1 | 75.2 |
| Metric3Dv2 | × | ✓ | 4.3 | **98.1** | - | - | **4.4** | **98.2** | 4.2 | 98.3 | **13.6** | **89.5** |
| GeoWizard | ✓ | ✓ | 5.2 | 96.6 | 6.1 | 95.3 | 9.7 | 92.1 | 6.4 | 96.1 | 29.7 | 79.2 |
| FE2E | ✓ | ✓ | 4.1 | 97.7 | 4.4 | 97.5 | 6.6 | 96.0 | **3.8** | **98.7** | 22.8 | 81.2 |
| Orchid | ✓ | ✓ | 5.7 | 96.9 | 6.3 | 95.8 | 7.7 | 94.4 | 7.3 | 96.9 | - | - |
| Diception | ✓ | ✓ | 7.2 | 93.9 | 7.5 | 93.8 | 7.5 | 94.5 | 5.3 | 96.7 | 24.3 | 74.1 |
| MUSE-g | ✓ | ✓ | 5.8 | 96.1 | 6.9 | 94.3 | 9.8 | 89.9 | 8.8 | 96.2 | 26.3 | 72.5 |
| MUSE-d | × | ✓ | 5.1 | 97.1 | 5.8 | 96.2 | 9.2 | 90.8 | 5.9 | 97.1 | 23.8 | 74.0 |

| Method | Gene- | MTL | NYUv2 | | ScanNet | | iBims-1 | | Sintel | | Oasis | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | rative | Model | Mean↓ | 11.25° ↑ | Mean↓ | 11.25° ↑ | Mean↓ | 11.25° ↑ | Mean↓ | 11.25° ↑ | Mean↓ | 11.25° ↑ |
| **Task 2: Surface Normal Estimation** | | | | | | | | | | | | |
| Marigold | ✓ | × | 17.1 | 58.5 | 16.1 | 62.3 | 16.6 | 68.0 | - | - | 23.5 | 28.0 |
| Lotus-g | ✓ | × | 16.9 | 59.1 | 15.3 | 64.0 | 17.5 | 66.1 | 35.2 | 19.9 | - | - |
| Lotus-d | × | × | 16.8 | 58.2 | 15.3 | 62.9 | 17.7 | 64.9 | 34.6 | 20.5 | - | - |
| Marigoldv1.1 | ✓ | × | 16.4 | 58.9 | 14.9 | 64.3 | 17.2 | 65.6 | - | - | **23.2** | 28.3 |
| Lotus-2 | × | × | 16.9 | 59.0 | 14.2 | 66.8 | 15.4 | 70.4 | **30.3** | **27.6** | - | - |
| Metric3Dv2 | × | ✓ | **13.2** | **66.2** | - | - | 19.6 | 69.7 | - | - | 23.4 | 28.5 |
| GeoWizard | ✓ | ✓ | 17.0 | 56.5 | 15.4 | 61.6 | **13.0** | 65.3 | - | - | - | - |
| FE2E | ✓ | ✓ | 16.2 | 59.6 | **13.8** | **67.2** | 15.1 | **70.6** | 31.2 | 22.3 | - | - |
| Orchid | ✓ | ✓ | 15.2 | 60.6 | 14.2 | 63.8 | 16.3 | 68.1 | 31.7 | 22.6 | - | - |
| Diception | ✓ | ✓ | 18.3 | 52.5 | 19.4 | 46.4 | - | - | - | - | - | - |
| MUSE-g | ✓ | ✓ | 17.1 | 58.0 | 15.7 | 62.7 | 17.4 | 65.8 | 35.9 | 19.0 | 26.2 | 26.6 |
| MUSE-d | × | ✓ | 16.6 | 58.6 | 15.4 | 61.3 | 16.7 | 67.3 | 35.2 | 18.6 | 23.4 | **30.5** |

models with single step inference, a total parameter size of only 0.95B, and a training set of only 59K. Detail computing resources comparison are provided in Supp. Materials Sec.C. Some failure cases of MUSE are shown in Sec.D. Moreover, experiments on different task numbers and token embedding methods are presented in Supp. Materials Sec.E

### 4.3 Manifold Decoupling Validation

To validate the Manifold Decoupling claim, we implement two methods to demonstrate that the zero-shot latent representations generated for multiple tasks are geometrically and semantically separable. A MUSE-d model trained with $t_{\text{rgb}} = 991, t_{\text{depth}} = 995, t_{\text{normal}} = 999$ is used as the predictor.

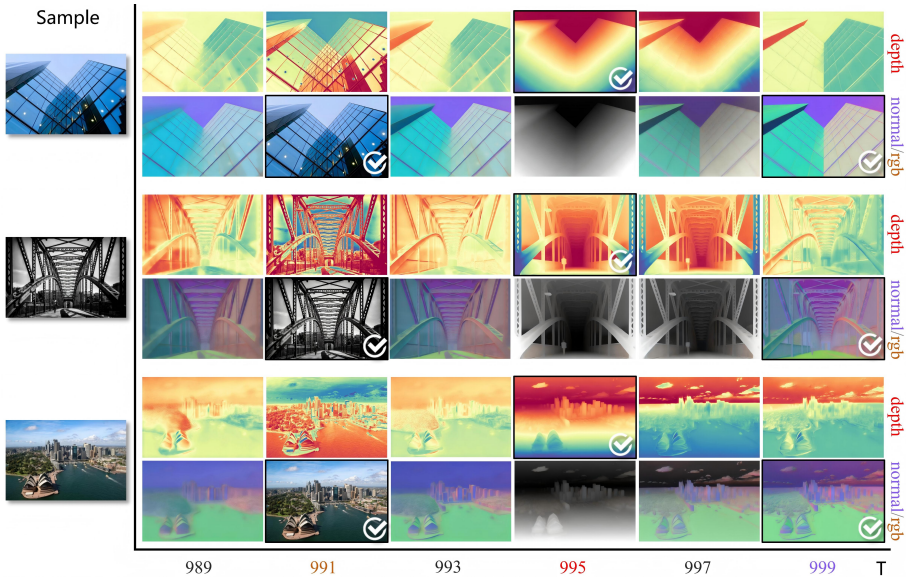


**Fig. 4:** Inference results at different timesteps by exactly the same model. The maps marked with a white checkmark are the expected outputs. Note that a sample generates only one prediction map at each timestep. The depth and normal/rgb images rows are two ways of visualizations, rendered by 1-channel and 3-channel respectively.

**Data Manifold of Latent Features.** We employ Isomap [54] to visualize the latent trajectories of diffusion models, where Isomap is a nonlinear dimensionality reduction technique that preserves global geodesic distances on a data manifold. Here we sample 200 images from coco2017 validation set [55]. For each image, we perform multiple forward passes using a sequence of different timesteps, extract the diffusion U-Net's final layer output latent. Then we apply multidimensional scaling to embed these latent vectors into a 2D plane, preserving their geodesic relationships.

The result is presented in the right of Figure 1. Latent features at timesteps where no steering function is applied are intertwined. While at timesteps 995 and 999, the latent vectors clearly segregate into two distinct clusters based on their corresponding task, demonstrating a clean separation in the geodesic feature space. This provides strong evidence that MUSE is not learning a single, entangled multi-task representation. Instead, it has successfully learned two separate data manifolds $\mathcal{M}_{\text{depth}}$ and $\mathcal{M}_{\text{normal}}$. The timestep embedding acts as the key, deterministically guiding the U-Net to project the input onto one of these two decoupled geometric structures. A quantitative metric table of Silhouette Score and Calinski-Harabasz is presented in Supp. Materials Sec.F.

**Intuitive Visualizations of Prediction Map.** To more intuitively illustrate the inference results of MUSE under different timesteps, Figure 4 shows the rendering of the model's prediction maps under two different visualization

ways. MUSE inference returns a 3-channel prediction map basically. For depth estimation, the channels are collapsed to a single scalar depth per pixel by taking the channelwise mean, and this single-channel map is then normalized and mapped to RGB using colormap. By contrast, normal estimation outputs are treated intrinsically as 3-channel vector fields. The predicted map is assumed to encode surface normals in the three channels and is directly converted to an RGB image for visualization. Naturally, it can observed that when the timestep is 991, MUSE's output is restored to the original image under the normal visualization, which is equivalent to performing an RGB image reconstruction task.

These visualization illustrates the internal mechanism of MUSE, providing a concrete validation for our theory.

### 4.4 Generality Across Architectures

To verify that the timestep steering mechanism is not merely an architectural artifact of the U-Net, we extend MUSE to the modern Diffusion Transformer (DiT) architecture [22]. Specifically, we replace the SDv2 U-Net with the SD3 MMDiT backbone, keeping the VAE frozen, and adapt the training paradigm to Conditional Flow Matching (CFM) [56]. Remarkably, the identical timestep steering protocol successfully decouples the depth and normal estimation tasks within the MMDiT backbone (Table 3, Supp. Materials Sec.G).

**Table 3: Quantitative metrics of MUSE based on DiT.** Variants using DiT as the backbone and flow matching as the optimization objective are implemented to verify the generalization of the MUSE mechanism. Experiments on 10 datasets show that the timestep steering is effective in Flow Matching, although the metrics are slightly lower than those of the U-Net version due to computational resource limitations.

| Method | Gene- | MTL | NYUv2 | | ScanNet | | KITTI | | ETH3D | | DIODE | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | rative | Model | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ | AbsRel↓ | $\delta_1$ ↑ |
| **Task 1: Affine-Invariant Depth Estimation** | | | | | | | | | | | | |
| MUSE-g-DiT | ✓ | ✓ | 8.6 | 92.4 | 10.1 | 88.9 | 11.6 | 86.8 | 9.0 | 92.4 | 23.4 | 71.9 |
| MUSE-d-DiT | × | ✓ | 8.0 | 93.7 | 8.4 | 92.3 | 10.7 | 87.3 | 10.8 | 92.0 | 24.1 | 72.0 |
| Method | Gene- | MTL | NYUv2 | | ScanNet | | iBims-1 | | Sintel | | Oasis | |
| | rative | Model | Mean↓ | 11.25° ↑ | Mean↓ | 11.25° ↑ | Mean↓ | 11.25° ↑ | Mean↓ | 11.25° ↑ | Mean↓ | 11.25° ↑ |
| **Task 2: Surface Normal Estimation** | | | | | | | | | | | | |
| MUSE-g-DiT | ✓ | ✓ | 20.4 | 51.3 | 22.0 | 45.9 | 19.1 | 61.0 | nan | 13.2 | nan | 11.3 |
| MUSE-d-DiT | × | ✓ | 19.6 | 50.5 | 19.7 | 52.1 | 19.5 | 59.8 | 38.4 | 13.7 | 34.5 | 24.5 |

Due to stringent computational constraints with only one V100-32GB GPU, this DiT variant necessitated aggressive optimization trade-offs, including FP16 mixed precision and 8-bit Adam optimizers. While these quantization effects result in slightly sub-optimal quantitative metrics compared to our fully-tuned U-Net baseline, the successful convergence and clear task decoupling serve as a

powerful proof-of-concept. It demonstrates that our proposed manifold decoupling via timestep steering is a fundamental, model-agnostic mechanism applicable to both classical diffusion and flow matching paradigms.

## 5 Discussion and Conclusion

This paper addresses the critical challenges of parameter inefficiency and negative transfer when adapting pre-trained diffusion models for unified visual perception. Rather than resorting to complex multi-head architectures, expert networks, heavy-parameter adapters or learnable task tokens, we introduce MUSE, a minimalist single-model multi-tasking paradigm. Our core conclusion establishes that the native, fixed timestep embedding — long considered a mere noise-level indicator — can be repurposed as a powerful, parameter-free semantic switch in one-step generation. By binding distinct timestep values to specific tasks, we successfully guide a single, full-parameter-shared model to perform decoupled geometric estimations. Extensive experiments validate MUSE's highly competitive performance across 10 diverse datasets, and demonstrate that this mechanism naturally generalizes across both classical U-Net and DiT architectures.

**Limitations on Semantic Segmentation.** While MUSE elegantly resolves inter-task conflicts for continuous geometric targets (e.g., depth and surface normals), our explorations reveal distinct boundaries when extending to discrete semantic tasks. For instance, we attempted to steer MUSE to concurrently generate NYUv2's 41-class semantic segmentation masks on the Hypersim dataset by formulating the discrete class IDs as continuous 3-channel RGB maps optimized via MSE loss. As anticipated, this resulted in sub-optimal quantitative metrics across the board. The semantic pixel accuracy dropped significantly, see Supp. Material Sec.H for detailed setups and results. We hypothesize this degradation is not a failure of the timestep steering mechanism itself, but rather a fundamental limitation of standard continuous VAEs in encoding and reconstructing highly discrete, non-smooth semantic ID maps without specialized heads or discrete latent spaces. Resolving this representation gap remains an open challenge for unified diffusion-based perception.

**Future Directions.** MUSE suggests that the path toward generalist models may not solely rely on brute-force scaling or stacking external expert modules. Instead, immense value lies in discovering more efficient ways to unlock and reshape the representational potential already dormant within a single foundation model. In the future, we plan to scale this paradigm to a wider set of semantic and intrinsic image decomposition tasks to probe the upper bounds of a single model's task capacity. Furthermore, while the current selection of orthogonal timesteps is empirical, exploring automated mechanisms to discover the most effective orthogonal task vectors within the time-embedding space presents a highly promising research direction.

## Acknowledgements

The research was supported by [Basic Research Center, Innovation Program of Chinese Academy of Agricultural Sciences (CAAS-BRC-SAE-2025-01)].

## References

1. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)
2. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
3. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
4. Lai, C.H., Song, Y., Kim, D., Mitsufuji, Y., Ermon, S.: The principles of diffusion models. arXiv preprint arXiv:2510.21890 (2025)
5. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in neural information processing systems **35**, 25278–25294 (2022)
6. Saxena, S., Kar, A., Norouzi, M., Fleet, D.J.: Monocular depth estimation using diffusion models. arXiv preprint arXiv:2302.14816 (2023)
7. Ji, Y., Chen, Z., Xie, E., Hong, L., Liu, X., Liu, Z., Lu, T., Li, Z., Luo, P.: Ddp: Diffusion model for dense visual prediction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21741–21752 (2023)
8. Duan, Y., Guo, X., Zhu, Z.: Diffusiondepth: Diffusion denoising approach for monocular depth estimation. In: European Conference on Computer Vision. pp. 432–449. Springer (2024)
9. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)
10. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., Chen, Y.C.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. arXiv preprint arXiv:2409.18124 (2024)
11. Krishnan, A., Yan, X., Casser, V., Kundu, A.: Orchid: Image latent diffusion for joint appearance and geometry generation. arXiv preprint arXiv:2501.13087 (2025)
12. He, J., Li, H., Sheng, M., Chen, Y.C.: Lotus-2: Advancing geometric dense prediction with powerful image generative model. arXiv preprint arXiv:2512.01030 (2025)
13. Wang, J., Lin, C., Guan, C., Nie, L., He, J., Li, H., Liao, K., Zhao, Y.: Jasmine: Harnessing diffusion prior for self-supervised depth estimation. arXiv preprint arXiv:2503.15905 (2025)
14. Ye, H., Xu, D.: Diffusionmtl: Learning multi-task denoising diffusion model from partially annotated data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 27960–27969 (2024)
15. Dong, L., Zhai, W., Zha, Z.J.: Unidense: Unleashing diffusion models with meta-routers for universal few-shot dense prediction. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 10525–10534 (2024)

16. Zhang, M., Song, G., Shi, X., Liu, Y., Li, H.: Three things we need to know about transferring stable diffusion to visual dense prediction tasks. In: European Conference on Computer Vision. pp. 128–145. Springer (2024)
17. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9970–9980 (2024)
18. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: European Conference on Computer Vision. pp. 241–258. Springer (2024)
19. Go, H., Lee, Y., Lee, S., Oh, S., Moon, H., Choi, S.: Addressing negative transfer in diffusion models. Advances in Neural Information Processing Systems **36**, 27199–27222 (2023)
20. Hang, T., Gu, S., Li, C., Bao, J., Chen, D., Hu, H., Geng, X., Guo, B.: Efficient diffusion training via min-snr weighting strategy. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7441–7451 (2023)
21. Javaloy, A., Valera, I.: Rotograd: Gradient homogenization in multitask learning. arXiv preprint arXiv:2103.02631 (2021)
22. Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4195–4205 (2023)
23. Lee, H.Y., Tseng, H.Y., Yang, M.H.: Exploiting diffusion prior for generalizable dense prediction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7861–7871 (2024)
24. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: What matters when repurposing diffusion models for general dense perception tasks? arXiv preprint arXiv:2403.06090 (2024)
25. Song, Z., Wang, Z., Li, B., Zhang, H., Zhu, R., Liu, L., Jiang, P.T., Zhang, T.: Depthmaster: Taming diffusion models for monocular depth estimation. arXiv preprint arXiv:2501.02576 (2025)
26. Ke, B., Qu, K., Wang, T., Metzger, N., Huang, S., Li, B., Obukhov, A., Schindler, K.: Marigold: Affordable adaptation of diffusion-based image generators for image analysis. arXiv preprint arXiv:2505.09358 (2025)
27. Frans, K., Hafner, D., Levine, S., Abbeel, P.: One step diffusion via shortcut models. arXiv preprint arXiv:2410.12557 (2024)
28. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. arXiv preprint arXiv:2505.13447 (2025)
29. Chen, Z., Zhou, M., Guo, J., Yuan, J., Ji, Y., Zhang, Y.: Steering one-step diffusion model with fidelity-rich decoder for fast image compression. arXiv preprint arXiv:2508.04979 (2025)
30. You, H., Liu, B., He, H.: Modular meanflow: Towards stable and scalable one-step generative modeling. arXiv preprint arXiv:2508.17426 (2025)
31. Park, B., Go, H., Kim, J.Y., Woo, S., Ham, S., Kim, C.: Switch diffusion transformer: Synergizing denoising tasks with sparse mixture-of-experts. In: European Conference on Computer Vision. pp. 461–477. Springer (2024)
32. Yang, Y., Jiang, P.T., Hou, Q., Zhang, H., Chen, J., Li, B.: Multi-task dense predictions via unleashing the power of diffusion. In: The Thirteenth International Conference on Learning Representations (2024)
33. Zhao, C., Sun, Y., Liu, M., Zheng, H., Zhu, M., Zhao, Z., Chen, H., He, T., Shen, C.: Diception: A generalist diffusion model for visual perceptual tasks. arXiv preprint arXiv:2502.17157 (2025)

34. Ma, Q., Ning, X., Liu, D., Niu, L., Zhang, L.: Decouple-then-merge: Finetune diffusion models as multi-task learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 23281–23291 (2025)
35. Crawshaw, M.: Multi-task learning with deep neural networks: A survey. arXiv preprint arXiv:2009.09796 (2020)
36. Seung, H.S., Lee, D.D.: The manifold ways of perception. science **290**(5500), 2268–2269 (2000)
37. Chung, H., Sim, B., Ryu, D., Ye, J.C.: Improving diffusion models for inverse problems using manifold constraints. Advances in Neural Information Processing Systems **35**, 25683–25696 (2022)
38. He, Y., Murata, N., Lai, C.H., Takida, Y., Uesaka, T., Kim, D., Liao, W.H., Mitsufuji, Y., Kolter, J.Z., Salakhutdinov, R., et al.: Manifold preserving guided diffusion. arXiv preprint arXiv:2311.16424 (2023)
39. Stanczuk, J., Batzolis, G., Deveney, T., Schönlieb, C.B.: Your diffusion model secretly knows the dimension of the data manifold. arXiv preprint arXiv:2212.12611 (2022)
40. Jacobsen, A.K., Gegenfurtner, J.M., Arvanitidis, G.: Staying on the manifold: Geometry-aware noise injection. arXiv preprint arXiv:2509.20201 (2025)
41. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
42. Stanczuk, J.P., Batzolis, G., Deveney, T., Schönlieb, C.B.: Diffusion models encode the intrinsic dimension of data manifolds. In: Forty-first International Conference on Machine Learning (2024)
43. Meilă, M., Zhang, H.: Manifold learning: What, how, and why. Annual Review of Statistics and Its Application **11**(1), 393–417 (2024)
44. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10912–10922 (2021)
45. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2 (2020)
46. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: European conference on computer vision. pp. 746–760. Springer (2012)
47. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013)
48. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5828–5839 (2017)
49. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3260–3269 (2017)
50. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019)
51. Koch, T., Liebel, L., Fraundorfer, F., Korner, M.: Evaluation of cnn-based single-image depth estimation methods. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
52. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: European conference on computer vision. pp. 611–625. Springer (2012)

53. Chen, W., Qian, S., Fan, D., Kojima, N., Hamilton, M., Deng, J.: Oasis: A large-scale dataset for single image 3d in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 679–688 (2020)
54. Tenenbaum, J.B., Silva, V.d., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. science **290**(5500), 2319–2323 (2000)
55. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
56. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)