ONE-STEP VIDEO DEPTH ESTIMATION VIA SELF-DISTILLATION

Wenqing Cui, Zhenyu Li, Jian Shi, Shariq Farooq Bhat & Peter Wonka
KAUST
{firstname.lastname}@kaust.edu.sa

ABSTRACT

Diffusion-based video depth estimation methods have recently set new benchmarks by leveraging rich generative priors learned from video synthesis, delivering exceptional depth accuracy and robust temporal consistency. However, the iterative nature of these models creates a computational bottleneck, hindering their utility in autonomous or dynamic environments that require real-time adaptation. To bridge this gap, we frame the efficiency-accuracy trade-off as a self-improvement challenge. We propose a two-stage self-distillation strategy. In the first stage, we distill a multi-step diffusion model into a one-step student by applying latent-space distillation to the Unet via score matching and latent gradient matching. In the second stage, we further distill the decoder using feature alignment and pixel-wise distillation losses. Our method achieves depth accuracy comparable to state-of-the-art multi-step video depth models, while reducing the denoising time by up to 3× and the decoding time by up to 20×.

1 INTRODUCTION

Video depth estimation is crucial in enabling numerous applications across computer vision, graphics, and robotics Chen et al. (2024); Feng et al. (2024); Guo et al. (2023); Holynski & Kopf (2018); Dong et al. (2021); Godard et al. (2019); Sun et al. (2021). Recent advances in diffusion-based generative models, such as *DepthCrafter* Hu et al. (2025), have significantly raised the bar for this task by leveraging rich generative priors learned from video synthesis Ke et al. (2024); Hu et al. (2025); Shao et al. (2024). These models demonstrate exceptional accuracy and temporal consistency even in dynamic and complex scenes.

However, the iterative denoising process inherent in diffusion models poses a critical computational bottleneck, leading to prohibitive inference times that hinder deployment in "Rapid or Frequent Inference Scenarios" (RFIS) Chen et al. (2025). Traditional solutions, such as training lightweight models from scratch, often compromise the generalization capabilities or require costly ground-truth data Xu et al. (2024).

In this paper, we frame this efficiency-accuracy trade-off not merely as a compression task, but as a **self-improvement challenge**. We propose a two-stage self-distillation strategy that enables the model to autonomously refine its inference efficiency by leveraging its own learned representations. To address the iterative denoising bottleneck, we first distill the multi-step *DepthCrafter* teacher into a one-step student model via latent-space distillation. This process successfully preserves the teacher's nuanced generative knowledge while collapsing the denoising process into a single forward pass without requiring external labeled data.

Beyond timestep distillation, we introduce an additional decoder distillation phase, replacing the heavy VAE decoder with a lightweight head trained through feature-wise and pixel-wise self-supervision. Our framework study both stochastic and deterministic paths to balance computational expressiveness and efficiency. Extensive experiments demonstrate that our distilled models match the accuracy of the original *DepthCrafter* while reducing UNet inference time by up to 3× and decoding time by up to 20×. By bridging the performance-efficiency divide through self-directed updates, our work advances the practicality of generative depth models in real-world, time-sensitive applications.

2 RELATED WORK

Diffusion-Based Depth Estimation. Diffusion models Song et al. (2022); Ho et al. (2020); Karras et al. (2022) are able to model complex data distributions with remarkable generative fidelity. Recent

advances such as *Stable Diffusion (SD)* Rombach et al. (2021) leverages billions of image-text pairs from the LAION-5B dataset Schuhmann et al. (2022), demonstrating the ability to achieve high-quality and diverse image generation with large-scale pretraining. Large-scale pretraining has also proven critical for advancing visual perception tasks Ravi et al. (2024); Ranftl et al. (2022); Yang et al. (2024b). Recent works Ke et al. (2024); Hu et al. (2025) repurpose diffusion models pretrained on large-scale data, achieving strong performance on various dense prediction tasks through lightweight fine-tuning. In monocular depth estimation, *Marigold* Ke et al. (2024) adapts a pretrained *SD* model by fine-tuning only its UNet component on synthetic depth data, allowing *Marigold* to leverage rich visual priors from the SD model. It achieves the state-of-the-art performances with strong generalization to in-the-wild images in a zero-shot setting. Extending this idea to video depth estimation, *DepthCrafter* Hu et al. (2025) fine-tunes *Stable Video Diffusion (SVD)* Blattmann et al. (2023) using a three-stage training pipeline on both synthetic and real-world datasets, without the need for camera poses or optical flow Chen et al. (2019); Kopf et al. (2021), making it applicable to diverse open-world videos. However, both *Marigold* and *DepthCrafter* rely on multi-step iterative denoising during inference, restricting their usability for real-world applications due to the high computational cost.

One-Step Diffusion via Distillation To reduce the computational burden brought by the iterative denoising within diffusion models. Recent research has explored strategies for one or few-step inference without sacrificing the generation quality. One line of work focuses on reformulating the diffusion process as a direct mapping from noise to data. *Rectified Flow* Liu & Song (2022) proposes learning a continuous flow that transforms noise into data in a single pass, effectively bypassing iterative refinement. *InstaFlow* Liu & Song (2023) extends this idea that demonstrates one-step generation is sufficient to produce high-quality text-to-image results. Another direction leverages knowledge distillation to compress multi-step diffusion processes into faster alternatives. Early works such as *Progressive Distillation* Salimans & Ho (2022) and *Guided Diffusion Distillation* Meng et al. (2023) train student models to mimic the behavior of multi-step teachers, achieving significant reductions in sampling steps. Recent works such as *Consistency Models* Song et al. (2023) and *Latent Consistency Models* Ho et al. (2023) extend this idea by enforcing the consistency between predictions at different noise levels, enabling high-quality synthesis with very few inference steps. While prior distillation efforts primarily target generative tasks, distilling a diffusion model for dense prediction tasks (*e.g.* depth estimation) remains underexplored.

3 Stochastic And Deterministic One-Step Distillation

We propose a one-step distillation framework for one-step video depth estimation for stochastic and deterministic diffusion models. In this section, we first elaborate preliminaries of our method. Next, we discuss the methodologies for stochastic and deterministic one-step distillation. Finally, we introduce a lightweight decoder head to achieve a faster inference speed. We show an illustration of our distillation framework in Fig. 1.

3.1 Preliminaries

Stable Video Diffusion In *SVD* Blattmann et al. (2023), a VAE encoder $\mathcal{E}$ first compresses the input video $\mathbf{X}$ into a latent representation $\mathbf{x}_0 = \mathcal{E}(\mathbf{X})$. By using the *Elucidated Diffusion Model (EDM)* Karras et al. (2022) framework, the diffusion process operates in the latent space by adding Gaussian noise with noise level σ_t, sampled from a log-normal distribution. Whilst training, the diffusion process is achieved by $\mathbf{x}_t = \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon},\ \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, where $\mathbf{x}_t$ represents the noisy latent at noise level σ_t at timestep t. The reverse diffusion denoiser $F(\mathbf{x}_t; \sigma_t; \mathbf{c})$ starts from a high-noise state $\sigma = \sigma_{\text{max}}$ that learns to gradually denoise the noisy latent $\mathbf{x}_t$ towards $\sigma = 0$ to recover the clean latent x_0. The final reconstructed video $y = \mathcal{D}(F(\mathrm{x}_t; \sigma_t; \mathbf{c}))$ can be obtained by the VAE decoder $\mathcal{D}$. *EDM*'s preconditioning strategy adds additional conditions to the denoiser, resulting in $F(\mathbf{x}_t; \sigma_t; \mathbf{c}) = c_{\text{skip}}(\sigma_t)\mathbf{x}_t + c_{\text{out}}(\sigma_t)U_\theta\left(c_{\text{in}}\mathbf{x}_t; c_{\text{noise}}(\sigma_t); \mathbf{c}\right)$, where c_{in}, c_{skip}, c_{out}, and c_{noise} are the preconditioning functions, U_θ is the learnable UNet that parametrized by θ. EDM employs a wide noise range. To stabilize training, an input-scaling condition c_{in} is applied on the noisy input $\mathbf{x}_t$:

$$\tilde{\mathbf{x}}_t = c_{\text{in}}\mathbf{x}_t = \frac{\mathbf{x}_t}{\sqrt{\sigma_t^2 + 1}} = \frac{\mathbf{x}_0 + \sigma_t\boldsymbol{\epsilon}}{\sqrt{\sigma_t^2 + 1}}. \tag{1}$$

When σ_t is large (*e.g.*, $\sigma_t = \sigma_{\text{max}}$), $\tilde{\mathbf{x}}_t$ becomes nearly indistinguishable from Gaussian noise, $\tilde{\mathbf{x}}_t \approx \epsilon$.

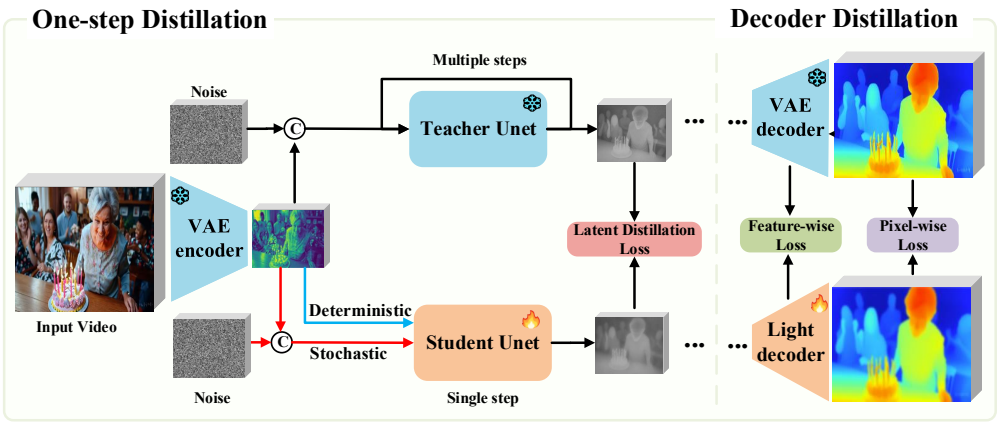


Figure 1: Our framework includes one-step distillation for the diffusion models (Sec. 3.2) and distillation of the VAE decoder to a lightweight decoder (Sec. 3.3). For one-step distillation, we denote the stochastic and deterministic forward pathways using the red and blue lines, respectively.

DepthCrafter Hu et al. (2025) reformulates depth estimation as a conditional generation task based on *SVD* and *EDM*, instead of predicting depth from RGB inputs. Given a ground-truth disparity video $\mathbf{D}$ and its corresponding RGB video $\mathbf{X}$, we obtain depth latent $\mathbf{r} = \mathcal{E}(\mathbf{X})$ and RGB latent $\mathbf{x}_0 = \mathcal{E}(\mathbf{D})$. With the same diffusion process as in *SVD*, we obtain noisy depth latent x_t. Therefore, the model learns to reconstruct clean depth latents $\mathbf{x}_0$ from the concatenation of $\mathbf{x}_t$ and $\mathbf{r}$, conditioned on the corresponding $\mathbf{r}$. For n-step inference, we may obtain the denoised latent as: $\hat{\mathbf{x}}_{0;n} = F_1(\cdot;\sigma_1) \circ \cdots \circ F_n(\cdot;\sigma_n)(\epsilon;\mathbf{c})$, where $F_n(\cdot;\sigma_n)$ represents a denoiser under a noise-level of σ_n.

3.2 Stochastic And Deterministic One-Step Distillation

A pretrained multi-step *DepthCrafter* is used as the *teacher model*. For any given input video $\mathbf{X}$, our goal is to distill a one-step *student model* to achieve similar performances against its multi-step teacher model without any labeled training data. This section describes our distillation methods for both stochastic (Sec. 3.2.1) and deterministic (Sec. 3.2.2) paths.

3.2.1 The Distillation For One-Step Diffusion

We fix the noise level to the maximum value $\sigma = \sigma_{\text{max}}$ for the student model during both training and inference. Referring to Eq. equation 1, the UNet input hereby becomes Gaussian noise. Essentially, we encourage the student model to learn a direct noise-to-video (noise-to-depth in this work) mapping, effectively collapsing the multi-step denoising process into a single forward pass. During the distillation, we only optimize the parameters of the UNet denoiser U_θ, which contains the majority of spatial visual knowledge as per Xu et al. (2024). Unlike prior works such as *DMD* Yin et al. (2024b) and *DMD2* Yin et al. (2024a), which distill in the pixel space, our method performs one-step distillation directly in the latent space with a *score matching distillation loss* and a *latent gradient distillation loss*, avoiding repeated VAE decoding during training.

Score Matching Distillation Loss. To align the denoising behavior of the student model with that of the teacher, we use a *score distillation loss* defined as: $\mathcal{L}_{\text{sm}} = \mathbb{E}\left[\left\|F(\mathbf{x}_t;\sigma_{\text{max}};c) - \hat{x}_{0;n}\right\|^2\right]$, where $F(\mathbf{x}_t;\sigma_{\text{max}};c)$ and $\hat{x}_{0;n}$ represent the latents from the one-step student model and n-step teacher model, respectively. This loss encourages the student model to mimic the teacher's denoising outputs in the latent space, allowing soft supervision from the teacher rather than supervised labels.

Latent Gradient Distillation Loss. We distill spatial gradients in the latent space. This imposes a stronger structural constraint on the student output. This can be formulated as: $\mathcal{L}_{\text{g}} = \frac{1}{M}\sum_{i=1}^{M}\left(|\nabla_x D_i| + |\nabla_y D_i|\right), \quad D = F(\mathbf{x}_t;\sigma_{\text{max}};c) - \hat{x}_{0;n}$, where D denotes the differences between the denoised latent outputs and we use D_i to represent the difference at spatial location i. ∇_x and ∇_y denotes gradient in x and y direction. Here, M represents the total number of spatial locations (pixels) in the latent feature map.

The total latent distillation loss becomes a weighted combination: $\mathcal{L}_{\text{latent}} = (1-\lambda)\mathcal{L}_{\text{sm}} + \lambda\mathcal{L}_{\text{g}}$.

3.2.2 A DETERMINISTIC ALTERNATIVE TO NOISE SAMPLING

Inspired by Lee et al. (2024); Xu et al. (2024), we further explore a deterministic variant of our diffusion pipeline by replacing the initial Gaussian noise with an RGB latent encoded by the VAE encoder. While stochastic sampling typically introduces variability in the generated outputs, especially for a deterministic prediction task such as depth prediction, this strategy effectively removes stochasticity from the sampling process.

Recall that the VAE is trained to produce latent representations with a KL divergence loss that approximates the standard normal distribution. The latent RGB features from the VAE encoder can be interpreted as a special case of noise that follows the same noise distribution. By using such features as deterministic initial inputs in place of random noise, we maintain generation quality while improving reproducibility and interpretability of the results.

3.3 A LIGHTWEIGHT DECODER HEAD

Compared to video generation tasks, depth videos normally contain less semantic information. A natural question to ask is that, *do we need that strong decoder for simpler depth problems?* Notably, VAE decoding is the primary runtime bottleneck for an *SVD* model as experimented in Tab. 2. Since deterministic pipeline produces noise-free features in the diffusion UNet as illustrated in Fig. 3, we can exploit UNet features to design and train a custom 'upsampling' decoder. We therefore replace the heavy VAE decoder with a simpler DPT-based architecture. We denote the DPT-based lightweight decoder as the student decoder and the temporal VAE decoder as the teacher decoder. For effectively distilling the knowledge from the teacher decoder to the student decoder, we employ *Feature-wise Distillation* (Sec. 3.3.1) and *Pixel-wise Distillation* (Sec. 3.3.2).

3.3.1 FEATURE-WISE DISTILLATION

We compare the pair-wise *affinity maps* of the features from the student and teacher decoders to align the feature between the student and teacher decoders Liu et al. (2019); Li et al. (2022). Let $m_s \in \mathbb{R}^{h\times w\times c_s}$ and $m_t \in \mathbb{R}^{h\times w\times c_t}$ denote the feature map produced by the student and teacher decoders, respectively. h and w are the spatial dimensions of feature maps, and c_s, c_t are the respective channel dimensions. We compute *affinity maps* based on pair-wise cosine similarity for each feature map. For each pair of spatial positions (i, j), the affinity scores for the student a^{s} and the teacher a^{t} are computed as $a_{\text{s}}^{ij} = \frac{\langle m_s(i), m_s(j)\rangle}{\|m_s(i)\|_2\|m_s(j)\|_2}$, $a_{\text{t}}^{ij} = \frac{\langle m_t(i), m_t(j)\rangle}{\|m_t(i)\|_2\|m_t(j)\|_2}$. The feature-wise distillation loss is defined as the mean squared error between the affinity maps of the student and teacher: $\mathcal{L}_{\text{feature}}(m_s, m_t) = \frac{1}{hw}\sum_i\sum_j \left(a_{\text{s}}^{ij} - a_{\text{t}}^{ij}\right)^2$, where the double summation runs over all spatial positions.

3.3.2 PIXEL-WISE DISTILLATION

Scale-Invariant Loss. We incorporate a *pixel-level scale-invariant loss* Farooq Bhat et al. (2021) to encourage the student decoder to produce depth outputs consistent with the teacher decoder at the pixel level. This loss penalizes the pixel-wise discrepancy between the teacher-student logarithmic predictions.

Let $d_s \in \mathbb{R}^{H\times W}$ and $d_t \in \mathbb{R}^{H\times W}$ denote the depth maps predicted by the student and teacher decoders, respectively. The *scale-invariant loss* $\mathcal{L}_{\text{silog}}$ is defined as: $\mathcal{L}_{\text{silog}}(d_s, d_t) = \alpha\sqrt{\frac{1}{T}\sum_{i,j}(g^{ij})^2 - \frac{\lambda}{T^2}\left(\sum_{i,j} g^{ij}\right)^2}$, where $g^{ij} = \log d_s^{ij} - \log d_t^{ij}$ is the per-pixel difference in log space, and T denotes the number of valid pixels used in the computation. We follow prior work Farooq Bhat et al. (2021) and set $\lambda = 0.85$ and $\alpha = 10$.

Gradient Matching Loss. We adopt the gradient matching loss to encourage the model to predict depth maps with sharp edges and fine-details, which is commonly used in training of depth estimation models. Gradient matching loss $\mathcal{L}_{\text{grad}}$ can be defined as: $\mathcal{L}_{\text{grad}}(d_s, d_t) =$

$\frac{1}{hw}\sum_i\sum_j\left(\left|\nabla_x K^{ij}\right|+\left|\nabla_y K^{ij}\right|\right)$, $K = d_s - d_t$ where K is the difference map between the student and teacher predictions, ∇_x and ∇_y donate the gradients in x and y directions, respectively.

Our overall loss function for distilling the lightweight decoder is defined as $\mathcal{L}_{\text{final}}(d_s, d_t, m_s, m_t) = \mathcal{L}_{\text{silog}}(d_s, d_t) + \lambda_1\mathcal{L}_{\text{feature}}(m_s, m_t) + \lambda_2\mathcal{L}_{\text{grad}}(d_s, d_t)$, where λ_1 and λ_2 are predefined coefficients for $\mathcal{L}_{\text{feature}}(m_s, m_t)$ and $\mathcal{L}_{\text{grad}}(d_s, d_t)$, respectively.

4 EXPERIMENTS

4.1 IMPLEMENTATION DETAILS

We adopt *DepthCrafter* Hu et al. (2025) as our base model, in which a fixed noise level of $\sigma = \sigma_{max}$ is used for the student model, where $\sigma_{\text{max}} = 700$ for the used `EulerDiscreteScheduler`. We operate in the disparity domain. The predicted disparity maps are normalized across frames to $[0, 1]$ range. *Virtual-KITTI* Cabon et al. (2020) and *Hypersim* Roberts et al. (2021) are used as our training datasets. We use Adam optimizer with a learning rate of 1×10^{-5}, batch size of one, and a gradient accumulation step size of 16. Four GPUs are used. More details can be found in our supplementary material.

The teacher and student models share the same UNet architecture. We use a frozen teacher model that performs a 5-step inference to guide the one-step student model for distillation. During the distillation for both stochastic and deterministic approaches, we optimize only the UNet within the student model with $\mathcal{L}_{\text{latent}}$ with $\lambda = 0.5$. Notably, we freeze the temporal layers of the student UNet model whilst training only the spatial layers. We use a DPT head Ranftl et al. (2021) with temporal modules from Video Depth Anything Chen et al. (2025) as the lightweight decoder. We keep the teacher model frozen while optimizing both the student UNet and the DPT head under $\mathcal{L}_{\text{final}}$ with $\lambda_1 = 0.1$ and $\lambda_2 = 2$. Each distillation takes $55k$ iterations, requiring approximately two days of distillation in total.

4.2 EVALUATION

Evaluation Datasets. We conduct a zero-shot evaluation study. We use scale-shift alignment to align the model prediction to the ground truth using the least-squares criterion Ranftl et al. (2022) and evaluate our model on three datasets spanning indoor and outdoor scenarios, with both static and dynamic scenes. **NYU-v2** Silberman et al. (2012) is a standard indoor benchmark for single-image depth estimation. We evaluate on its official test set, which contains 654 RGB-D images. **KITTI** Geiger et al. (2013) is a street-scene outdoor dataset collected for autonomous driving. It provides sparse LiDAR-based ground-truth depths. We use the official validation split, which includes 13 scenes. To evaluate temporal consistency and generalization on long-range sequences, we extract 13 continuous video sequences, each consisting of 110 frames. **Bonn** Palazzolo et al. (2019) is a dataset of dynamic indoor scenes captured with an RGB-D camera. We select five representative video sequences from this dataset, each with a length of 110 frames, following the criterion of *DepthCrafter*.

Competing Methods. We use *DepthCrafter* as our baseline model and compare with representative image depth estimation methods (*e.g. Depth-Anything-V2* Yang et al. (2024c), *Marigold* Ke et al. (2024)) and video depth estimation methods such as *NVDS* Wang et al. (2023) and *ChronoDepth* Shao et al. (2024). We adopt the Depth-Anything-v2-large model and LCM version of Marigold with an ensemble size set to 5.

Evaluation Metrics. We majorly evaluate our model using the two accuracy metrics. We report Absolute Relative Error (AbsRel) and δ_1 accuracy to assess per-pixel depth prediction quality. In addition, we measure temporal consistency and boundary accuracy in our ablation studies.

4.3 RESULTS

Quantitative Results. We present *zero-shot* evaluation results in Tab. 1 and inference time in Tab. 2. As shown in Tab. 1, our distilled models outperform the original 1-step version of *DepthCrafter* and achieve comparable performances to the original 5-step *DepthCrafter* with only one-step denoising. As indicated in Tab. 4, this reduces the denoising time by approximately three times compared to the

5-step version. By distilling a lightweight DPT head, we achieve a 20× reduction in decoding time, while still achieving performance comparable to the 5-step version. The results demonstrate that our distillation framework effectively maintains accuracy while accelerating inference.

Table 1: Zero-shot evaluation on depth estimation benchmarks with our distilled one-step models. Red and blue colored numbers represent the best and the second best results, respectively.

Method		Video	Inference Steps	ms/frame	NYU-v2		KITTI		Bonn	
					AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Feed-forward Models.										
Depth-Anything-V2 †		×	–	180.46	0.043	0.978	0.140	0.804	0.106	0.921
NVDS†		✓	–	346.20	0.151	0.780	0.253	0.588	0.167	0.766
Diffusion-based Models.										
Marigold †		×	–	1070.29	0.070	0.946	0.149	0.796	0.091	0.931
ChronoDepth †		✓	5	308.98	0.073	0.941	0.167	0.759	0.100	0.911
Baseline	DepthCrafter (5-step)†	✓	5	450.97	0.072	0.948	0.104	0.896	0.071	0.972
	DepthCrafter (1-step)†	✓	1	317.86	0.082	0.935	0.138	0.812	0.084	0.954
Ours	Stochastic (VAE)	✓	1	317.86	0.080	0.939	0.102	0.898	0.065	0.966
	Deterministic (VAE)	✓	1	317.86	0.076	0.942	0.107	0.881	0.071	0.966
	Deterministic (DPT)	✓	1	147.40	0.082	0.931	0.110	0.876	0.075	0.961

†: Reported results from *DepthCrafter*.

Table 2: Inference time per frame (ms). Compared with diffusion-based depth models, our method achieves the fastest inference speed.

Method	Encoding	Denoising	Decoding	All
Marigold	256.40	114.53	699.36	1070.29
DepthCrafter (5-step)	90.24	178.80	181.66	450.7
Our (Stochastic + VAE)	90.24	49.57	178.04	317.85
Our (Deterministic + VAE)	90.24	49.57	178.04	317.85
Our (Deterministic + DPT)	90.24	49.57	**8.33**	**148.14**

Qualitative Results. Fig. 2 presents qualitative comparisons between our distilled one-step model and the baseline. We select representative samples from KITTI, Sintel, and NYU-v2 datasets to cover a diverse set of outdoor, indoor, and synthetic scenarios. As shown in the KITTI example, our method better preserves fine-grained structures, such as the edges and textures of roadside bushes, compared to the 1-step baseline. The Sintel and NYU-v2 samples similarly highlight improvements in structural sharpness and boundary consistency. These results demonstrate that our one-step distillation framework successfully transfers fine-detail knowledge from the multi-step teacher to the student model, yielding sharper depth predictions under various real and synthetic conditions.

5 Ablations

In this section, we perform ablation studies to show the effectiveness of our distillation methods. More ablations and experiments can be found in our supplementary material.

The one-step distillation. We conduct ablation experiments on the *Virtual-KITTI* dataset to evaluate the effectiveness of our one-step distillation components in Tab. 3. We start from a baseline student trained for multi-step denoising using score matching, but use only one step during inference (*Exp. A*). We observe that one-step distillation using score matching (*Exp. B*) significantly improves performance across all metrics. Introducing Latent-Gradient loss (*Exp. C*) and freezing temporal layers (*Exp. D*) further improves the error and enhances boundary sharpness. Finally, using deterministic pipeline (*Exp. E*) achieves the best overall results, with the lowest AbsRel (0.176), highest δ_1 (0.771), and most accurate edge alignment (Boundary-F1: 0.273).

Ablation study on decoder distillation. We also ablate the design of our decoder distillation loss(Tab. 4). Starting from a baseline using only the scale-invariant (SILog) loss (*Exp. A*), we observe that adding affinity map loss (*Exp. B*) improves performance, particularly in δ_1. Further incorporating gradient-matching (GM) loss (*Exp. C*) leads to the best results, reducing AbsRel to

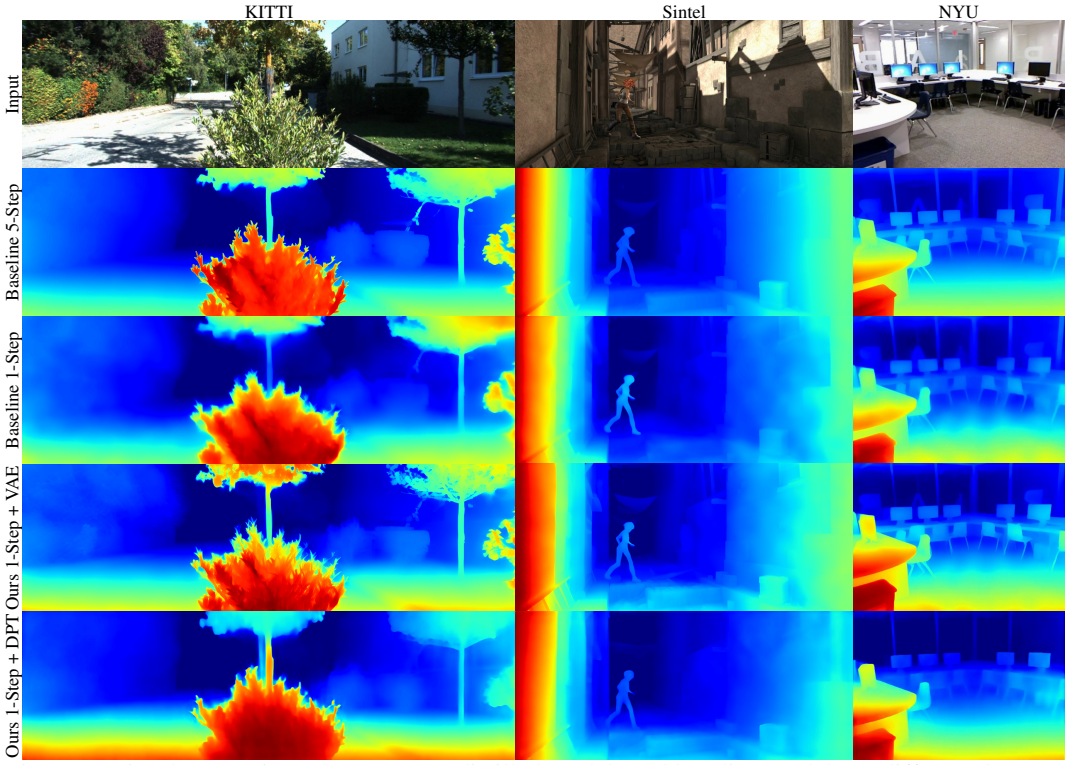


Figure 2: Visual comparisons between our distilled models and baseline models on different datasets. Our model presents a better performance on fine-details, especially when compared against the one-step baseline model.

Table 3: Ablation study of one-step distillation.

Exp.	One-step Forward	Latent-Gradient Loss	Freeze Temporal Layers	Deterministic	Virtual-KITTI			
					AbsRel↓	δ_1↑	TAE↓	Boundary-F1↑
A					0.245	0.670	0.201	0.191
B	✓				0.194	0.748	0.204	0.198
C	✓	✓			0.180	0.758	0.207	0.204
D	✓	✓	✓		0.190	0.766	0.187	0.254
E	✓	✓	✓	✓	**0.176**	**0.771**	**0.182**	**0.273**

0.185 and increasing δ_1 to 0.767. This confirms that both structural alignment (via affinity maps) and edge-aware supervision (via GM loss) are complementary and beneficial for training a compact yet effective decoder.

Table 4: Ablation study of decoder distillation.

Exp.	SILog loss	Affinity map loss	GM loss	Virtual-KITTI			
				AbsRel↓	δ_1↑	TAE↓	Boundary-F1↑
A	✓			0.198	0.756	**0.190**	0.097
B	✓	✓		0.192	0.763	0.193	0.106
C	✓	✓	✓	**0.185**	**0.767**	0.192	**0.116**

Temporal consistency. To evaluate the temporal stability of depth predictions across consecutive video frames, we adopt the Temporal Alignment Error (TAE) metric introduced in Depth Any Video Yang et al. (2024a). This metric quantifies the reprojection error between depth maps of adjacent frames using known camera poses.

Given a sequence of N frames, let d^k denote the predicted depth of the k-th frame, and p^k denote the transformation matrix (pose) from the k-th to the $(k+1)$-th frame. The TAE metric is computed as: $TAE = \frac{1}{2(N-1)} \sum_{k=1}^{N-1} \left[AbsRel\left(f(d^k, p^k), d^{k+1}\right) + AbsRel\left(f(d^{k+1}, p_{-1}^{k+1}), d^k\right) \right]$, where $f(\cdot, p)$ denotes the projection function that reprojects the depth map into the next (or previous) frame using

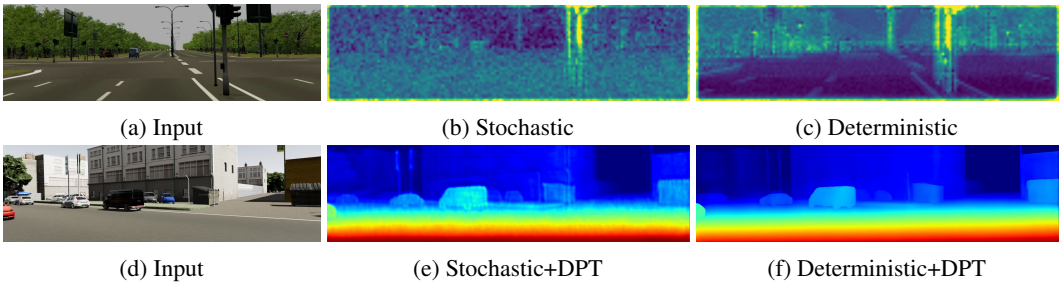
the transformation matrix p. p_{-1}^{k+1} is the inverse matrix for inverse projection. A lower TAE indicates better temporal coherence across the predicted depth sequence.

Boundary accuracy. We adopt the boundary F1 score from *DepthPro* Bochkovskii et al. (2025) to assess the accuracy of depth discontinuities. This metric compares depth ratios between neighboring pixels to detect occlusion boundaries and evaluates the overlap between predicted and ground-truth contours in terms of precision and recall. Following Bochkovskii et al. (2025), we average the F1 score over thresholds $t = 5$ to 25, with higher weights on larger thresholds. Since real-world data often contains noisy boundaries, we compute this metric on the synthetic *Sintel* Butler et al. (2012) dataset with dense ground truth.

Table 5: Boundary F1 score performances comparison on *Sintel*.

	DepthCrafter	Ours (Sto.)	Ours (Det.)	Ours (Det. + DPT)
Boundary-F1↑	0.218	0.216	**0.228**	0.173

Intermediate features for training DPT heads. We adopt a temporal-DPT Chen et al. (2025) decoder in this work. Different from the VAE decoder that only require denoised latents from UNet, a DPT-like structure commonly require a multi-scale feature pyramid. Interestingly, as shown in Fig. 3, though the last layer output latent features are always clean, the intermediate UNet features trained from a stochastic model are noisy due to the UNet skip-connection strategy.



(a) Input (b) Stochastic (c) Deterministic

(d) Input (e) Stochastic+DPT (f) Deterministic+DPT

Figure 3: Visualization of intermediate UNet features trained in stochastic (b) and deterministic (c) manners. The deterministic model yields cleaner features, resulting in sharper and clearer predictions in (f) compared to (e).

6 Conclusion

In this work, we presented a self-distillation framework that frames the efficiency-accuracy trade-off in diffusion-based video depth estimation as a **self-improvement challenge**. We demonstrated that the computational complexity of multi-step denoising can be effectively mitigated by enabling the model to internalize its generative knowledge into a single-step forward pass. By utilizing latent score matching and latent gradient matching, the student model successfully preserves the structural nuances of the original diffusion priors while significantly optimizing its own inference logic.

Our results show that the dual-stage refinement—addressing both the denoising UNet and the VAE decoder—allows the model to achieve a 3× speedup in denoising and a 20× speedup in decoding. This efficiency gain is achieved without compromising the zero-shot generalization and temporal consistency that are characteristic of powerful multi-step teachers. This progress moves generative depth estimation from a computationally heavy vision to a practical system capable of rapid execution in dynamic environments.

Ultimately, this work serves as a principled step toward self-improving AI systems that can autonomously optimize their operational system design. While we focused on timestep distillation within a fixed architecture, the success of this approach provides a foundation for future recursive improvements. Our next goal is to explore cross-architecture and latent-space evolution to further bridge the gap between high-fidelity generative modeling and the strict constraints of real-time, interactive applications.

ACKNOWLEDGMENT

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST) – Center of Excellence for Generative AI, under award number 5940 and a gift from Google.

REFERENCES

Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani, and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. URL https://arxiv.org/abs/2311.15127.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second, 2025. URL https://arxiv.org/abs/2410.02073.

D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In A. Fitzgibbon et al. (Eds.) (ed.), *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pp. 611–625. Springer-Verlag, October 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2, 2020.

Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. *arXiv:2501.12375*, 2025.

Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video diffusion models with motion prior and reward feedback learning, 2024. URL https://arxiv.org/abs/2305.13840.

Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera, 2019. URL https://arxiv.org/abs/1907.05820.

Xingshuai Dong, Matthew A. Garratt, Sreenatha G. Anavatti, and Hussein A. Abbass. Towards real-time monocular depth estimation for robotics: A survey, 2021. URL https://arxiv.org/abs/2111.08600.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4008–4017. IEEE, June 2021. doi: 10.1109/cvpr46437.2021.00400. URL http://dx.doi.org/10.1109/CVPR46437.2021.00400.

Ruoyu Feng, Wenming Weng, Yanhui Wang, Yuhui Yuan, Jianmin Bao, Chong Luo, Zhibo Chen, and Baining Guo. Ccedit: Creative and controllable video editing via diffusion models, 2024. URL https://arxiv.org/abs/2309.16496.

Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The international journal of robotics research*, 32(11):1231–1237, 2013.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation, 2019. URL https://arxiv.org/abs/1806.01260.

Yuwei Guo, Ceyuan Yang, Anyi Rao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Sparsectrl: Adding sparse controls to text-to-video diffusion models, 2023. URL https://arxiv.org/abs/2311.16933.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Jonathan Ho, Tim Salimans, William Chan, Mohammad Norouzi, David J Fleet, et al. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2305.01874*, 2023.

Aleksander Holynski and Johannes Kopf. Fast depth densification for occlusion-aware augmented reality. *ACM Trans. Graph.*, 37(6), December 2018. ISSN 0730-0301. doi: 10.1145/3272127.3275083. URL https://doi.org/10.1145/3272127.3275083.

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 *CVPR*, 2025.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models, 2022. URL https://arxiv.org/abs/2206.00364.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation, 2021. URL https://arxiv.org/abs/2012.05901.

Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction, 2024. URL https://arxiv.org/abs/2311.18832.

Zhenyu Li, Zehui Chen, Jialei Xu, Xianming Liu, and Junjun Jiang. Litedepth: digging into fast and accurate depth estimation on mobile devices. In *European Conference on Computer Vision*, pp. 507–523. Springer, 2022.

Ming Liu and Yang Song. Flow straight and fast: Learning to generate and transfer data with rectified flow. *arXiv preprint arXiv:2209.14687*, 2022.

Ming Liu and Yang Song. Instaflow: One step is enough for high-quality diffusion-based text-to-image generation. *arXiv preprint arXiv:2311.00391*, 2023.

Yifan Liu, Ke Chen, Chris Liu, Zengchang Qin, Zhenbo Luo, and Jingdong Wang. Structured knowledge distillation for semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2604–2613, 2019.

Chenlin Meng, Yang Song, and Stefano Ermon. On distillation of guided diffusion models. *arXiv preprint arXiv:2303.01539*, 2023.

Emanuele Palazzolo, Jens Behley, Philipp Lottes, Philippe Giguère, and Cyrill Stachniss. Refusion: 3d reconstruction in dynamic environments for rgb-d cameras exploiting residuals, 2019. URL https://arxiv.org/abs/1905.02082.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021.

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*, 44(3), 2022.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. URL https://arxiv.org/abs/2408.00714.

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 *International Conference on Computer Vision (ICCV) 2021*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. URL https://arxiv.org/abs/2210.08402.

Jiahao Shao, Yuanbo Yang, Hongyu Zhou, Youmin Zhang, Yujun Shen, Vitor Guizilini, Yue Wang, Matteo Poggi, and Yiyi Liao. Learning temporally consistent video depth from video diffusion priors, 2024. URL https://arxiv.org/abs/2406.01493.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European Conference on Computer Vision (ECCV)*, 2012.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Yang Song, Chenlin Meng, and Stefano Ermon. Consistency models. *arXiv preprint arXiv:2303.00752*, 2023.

Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video, 2021. URL https://arxiv.org/abs/2104.00681.

Yiran Wang, Min Shi, Jiaqi Li, Zihao Huang, Zhiguo Cao, Jianming Zhang, Ke Xian, and Guosheng Lin. Neural video depth stabilizer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9466–9476, October 2023.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? *arXiv preprint arXiv:2403.06090*, 2024.

Honghui Yang, Di Huang, Wei Yin, Chunhua Shen, Haifeng Liu, Xiaofei He, Binbin Lin, Wanli Ouyang, and Tong He. Depth any video with scalable synthetic data. *arXiv preprint arXiv:2410.10815*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024b.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024c.

Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024a.

Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024b.