DFM: Steering Marigold with DINOv2 Priors for Monocular Depth Estimation

Wei Shi, Zixuan Wu, Yucheng Zhu, Chenglan Huang, Tingting Han*, Min Tan*
School of Computer Science, Hangzhou Dianzi University, Hangzhou, China
weishi@hdu.edu.cn, ttinghan@hdu.edu.cn, tanmin@hdu.edu.cn

***Abstract*—Monocular depth estimation from a single image remains challenging, especially for sharp boundaries and fine-grained textures. To address this, we propose DFM (DINOv2-Fused Marigold), a framework that integrates DINOv2 features into the Stable-Diffusion-based Marigold pipeline via a hierarchical ViT encoder. Unlike conventional feature fusion, our encoder selectively amplifies informative channels while suppressing redundant ones, yielding a discriminative, geometry-aware representation that integrates global structural priors with precise local geometry. This design results in sharper boundaries, richer textures, and improved depth accuracy. Analysis of internal activations indicates that channel-wise co-variation in select DINOv2 channels sparsifies noisy responses while amplifying salient geometric structures, providing insight into the improved robustness and stability observed in our evaluations. Extensive experiments demonstrate that DFM fully outperforms the Marigold baseline on challenging outdoor benchmarks, while exhibiting competitive performance on indoor benchmarks, and simultaneously achieving improved training stability.**

***Index Terms*—DINOv2, Marigold, feature complementarity, multi-scale feature enhancement**

I. Introduction

Monocular depth estimation aims to reconstruct the three-dimensional geometry of a scene from a single image by predicting pixel-wise depth. Early approaches [1] were constrained by scene-specific training data and limited model capacity, resulting in suboptimal generalization. To overcome these limitations, subsequent studies explored a range of directions, including multi-scale feature fusion [2], adversarial and unsupervised learning [3], [4], vision–language model fine-tuning [5], and hybrid Transformer–CNN architectures [6], [7].

More recently, diffusion-based models [8] have shown strong promise for monocular depth estimation. Methods such as Marigold [9], EcoDepth [10], PrimeDepth [11], D4RD [12], and DiffusionDepth [13] leverage the powerful generative priors of Stable Diffusion [14], leading to improved structural preservation, zero-shot generalization, and robustness. Despite these advances, a fundamental limitation persists: the diffusion U-Net backbone lacks explicit geometric modeling, which restricts its ability to accurately recover precise sharp object boundaries and fine-grained geometric details.

Prior work indicates that incorporating complementary geometric cues can help mitigate this limitation. Existing studies have demonstrated that combining global structural priors with local geometric information yields more complete scene representations [2], [7]. Moreover, cross-resolution and cross-modality approaches, including BTS [15], HR-Depth [16], AdaBins [17], and BinsFormer [18], consistently show that global–local and semantic–geometric cues are intrinsically complementary for recovering fine-grained depth.

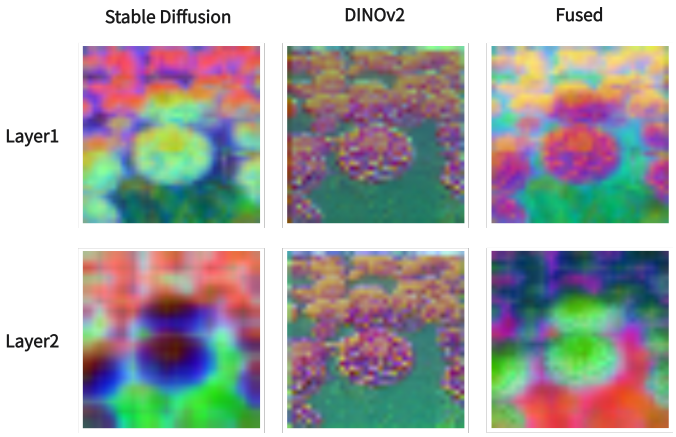

Fig. 1. Qualitative comparison of feature maps from Stable Diffusion's U-Net (left), DINOv2 (center), and the linearly fused representation (right).

In this context, recent evidence suggests that features from Stable Diffusion and DINOv2 lie in low-redundancy semantic subspaces and provide complementary global and local information [19]. Building on this observation, we analyze these characteristics empirically in our experiments, as shown in Fig. 1. Specifically, DINOv2 features are highly sensitive to local geometric variations and object boundaries, whereas features from the Stable Diffusion U-Net encoder primarily preserve global structural coherence and scene layout. Importantly, we find that even a simple linear fusion of these two feature streams can consistently yield representations with clearer contours and richer structural details.

Based on this, we propose DFM (DINOv2-Fused Marigold), a novel monocular depth estimation framework that explicitly integrates geometry-sensitive representations into a diffusion-based pipeline for improved and robust depth prediction. Built upon the Marigold framework, DFM introduces a Hierarchical ViT Encoder that extracts discriminative geometric features from DINOv2 and seamlessly integrates them into the diffusion U-Net backbone [20]. The hierarchical design enables

*Corresponding Author.

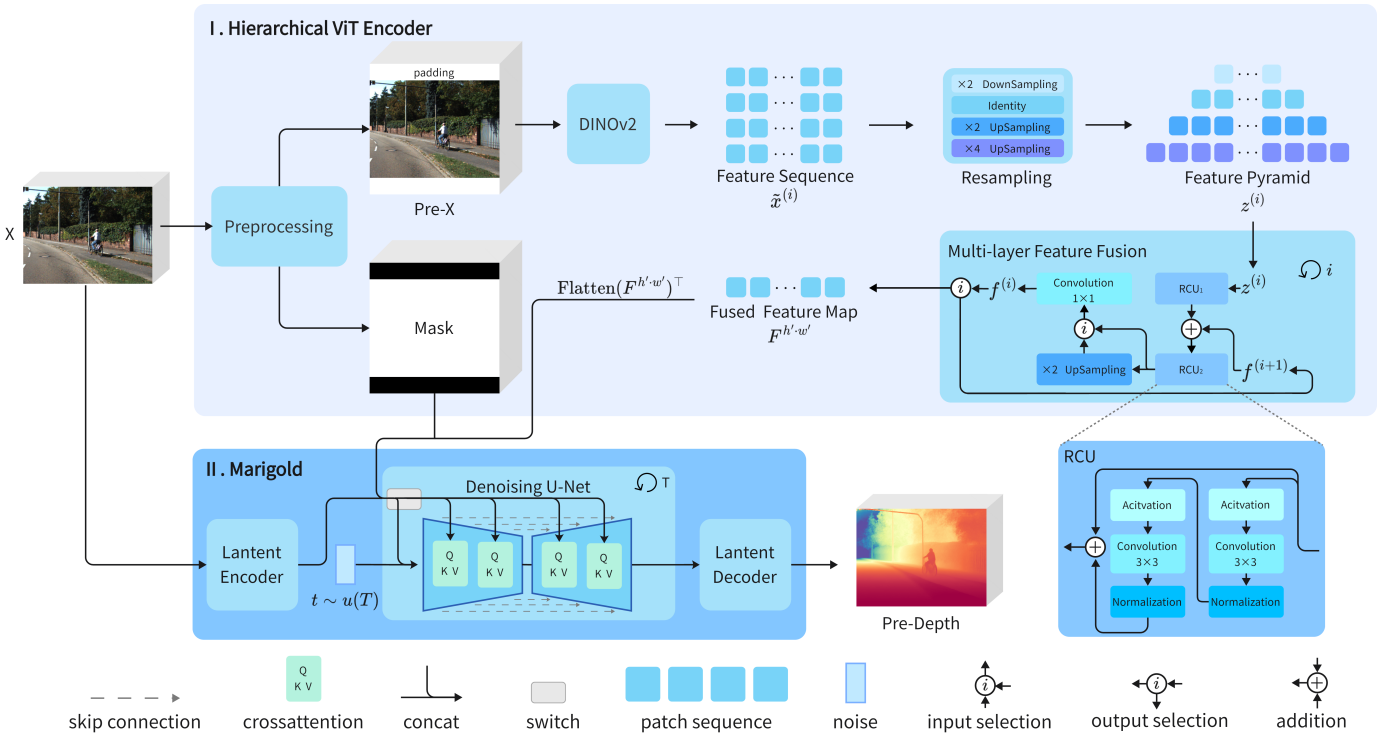

Fig. 2. Overview of the proposed DFM framework. The input image X is processed by a Hierarchical ViT Encoder to generate a padded image (Pre-X) and a mask map. DINOv2 extracts a feature sequence from the Pre-X, which is then resampled to construct a feature pyramid. This pyramid undergoes multi-layer fusion to produce a hybrid feature map F (where, during the first fusion stage, $f^{(i+1)} = \text{up}(z^{(i+1)})$, and the final feature fusion does not involve upsampling). Subsequently, F is processed and fed, together with the mask map, into the cross-attention layer of Marigold, where spatial features guide the image generation process.

TABLE I

FEATURE STATISTICS BEFORE AND AFTER FUSION. WE COMPARE FUSED FEATURES (FUSED) WITH RAW DINOV2 FEATURES FROM FOUR BACKBONE LAYERS (37 × 37 RESOLUTION, 1024 CHANNELS). MEAN AND MAX DENOTE THE AVERAGE AND MAXIMUM SPATIAL STANDARD DEVIATION ACROSS CHANNELS, WHILE MASKING INDICATES THE CHANNEL MASKING RATIO. WE REPORT THE MIN–MAX RANGE OVER FIVE DATASETS (NYU-V2, KITTI, ETH3D, SCANNET, DIODE).

Feature Name	Mean(10^{-3})	Max(10^{-3})	Masking
Fused features	1.86-1.96	102.25-108.90	0.77-0.78
Layer4	0.83-2.17	1.59-5.41	–
Layer11	0.71-1.87	1.24-3.56	–
Layer17	0.65-1.70	1.09-3.40	–
Layer23	0.64-1.65	1.04-3.22	–

effective multi-scale fusion of geometric information, allowing the model to jointly exploit the global structural priors of Stable Diffusion and the precise local cues provided by DINOv2. As a result, DFM produces sharper object boundaries, richer surface textures, and more accurate depth predictions. Project Page: https://github.com/DIGR-cyber/DFM.

Our main contributions are summarized as follows:

- We investigate the complementary strengths of Stable Diffusion and DINOv2 for monocular depth estimation, combining global structure with precise local geometry.
- We propose a Hierarchical ViT Encoder to efficiently fuse DINOv2 features into the U-Net backbone, enabling stable and effective feature integration.
- Our method significantly outperforms baseline method on outdoor benchmarks, with strong generalization and improved training stability across diverse scenarios.

II. METHOD

Fig. 2 illustrates the overall architecture of the proposed DFM framework. Given an input image, DFM extracts multi-scale geometry-aware features from DINOv2 and progressively fuses them through a hierarchical pyramid fusion scheme to form a unified representation. The fused features are then aligned and injected into the diffusion U-Net backbone of Marigold as conditional guidance during the denoising process. This design enables DFM to incorporate geometry-sensitive information into the diffusion-based depth estimation pipeline while preserving the original Marigold objective.

A. Hierarchical ViT Encoder

Feature Extraction. The input image is first preprocessed into a padded image with a valid feature mask. A frozen DINOv2 backbone is then employed to extract token-level feature sequences $\tilde{x}^{(i)}$ from selected layers j:

$$\tilde{x}^{(i)} = \text{Backbone}(\text{layer } j). \tag{1}$$

Each sequence $\tilde{x}^{(i)}$ is rearranged into a spatial feature map $z^{(i)}$. All feature maps are aligned to a common feature

TABLE II

QUANTITATIVE COMPARISON ON NYUV2, KITTI, ETH3D, SCANNET, AND DIODE BENCHMARKS. ABSOLUTE RELATIVE ERROR (ABSREL, ↓) AND ACCURACY (δ_1, ↑) ARE REPORTED. "*" DENOTES THE RE-IMPLEMENTED BASELINE RESULTS OBTAINED UNDER OUR UNIFIED EVALUATION PIPELINE. "(W/ ENSEMBLE)" SIGNIFIES THE USE OF TEST-TIME ENSEMBLING. UNLESS OTHERWISE SPECIFIED, ALL RESULTS ARE OBTAINED WITH THE DEFAULT MARIGOLD TEST-TIME CONFIGURATION OF 50 DENOISING STEPS AND AN ENSEMBLE SIZE OF 10. THE BEST RESULT FOR EACH METRIC IS HIGHLIGHTED IN BOLD.

Method	NYUv2		KITTI		ETH3D		ScanNet		DIODE	
	AbsRel ↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel ↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Marigold*(w/o ensemble)	6.32	95.41	10.66	89.66	6.98	95.75	7.19	93.10	30.86	77.28
Marigold*(w/ensemble)	5.78	96.07	10.03	90.90	6.46	96.09	**6.64**	**93.80**	30.21	77.54
Ours(w/o ensemble)	6.05	95.65	10.22	90.71	6.84	95.86	7.24	93.18	30.68	77.33
Ours(w/ensemble)	**5.58**	**96.36**	**9.82**	**92.15**	**6.27**	**96.32**	6.69	93.69	**29.92**	**77.89**

dimensionality while preserving distinct spatial resolutions r1, r2, r3, r4, forming a bottom-up multi-scale representation.

Multi-layer Pyramid Fusion. To integrate $z^{(i)}$ into a unified representation, we adopt a pyramid-style fusion strategy [21] that performs feature fusion in a top-down manner, progressively propagating information from deeper layers to shallower ones via a hierarchical fusion mechanism.

Specifically, the fusion process starts from the deepest layer and progressively proceeds toward shallower layers as the layer index i decreases. Before each fusion step, the fused feature map from the deeper level (i.e., the output of RCU2) is first upsampled to match the spatial resolution of the current level. At level i, the current feature map $z^{(i)}$ is first processed by RCU1 for feature adaptation. Subsequently, the output of RCU1 is combined with the upsampled fused feature from the previous deeper level, and the resulting feature is further refined by RCU2 in a sequential manner to obtain the fused representation at this level.

$$f^{(i)} = \mathrm{RCU}_2\Big(\mathrm{RCU}_1\Big(z^{(i)}\Big) + f^{(i+1)}\Big), \quad 1 \le i \qquad (2)$$

The final representation $f^{(0)}$ is subsequently aligned and projected, and a hybrid feature map F is obtained through downsampling and channel dimension expansion for improved feature compatibility. This feature map is then fed into the cross-attention layers of the U-Net for depth generation.

We then flatten the fused feature map F into a sequence T:

$$T = \mathrm{Flatten}(F)^\top \in \mathbb{R}^{B\times N\times C_d}, \quad N = h' \cdot w', \qquad (3)$$

where h' and w' are the latent spatial resolutions and N is the total number of tokens. Each token in T encodes a semantic vector with positional information. The sequence T is supplied as the Key/Value tensors for cross-attention [6], allowing the U-Net to incorporate image semantics at every denoising step.

B. Functional Role of the Hierarchical ViT Encoder

To elucidate the role of the proposed Hierarchical ViT Encoder, we analyze its internal activations and quantitatively examine how it reshapes input features at the channel level, from a representation learning perspective, aiming to explain why DFM achieves superior depth estimation performance. While the proposed metrics are not explicitly optimized during training, they provide an interpretable and principled perspective for systematic and fine-grained analysis of the emergent behavior induced by the learned encoder.

The results show that the encoder functions as a *channel selector*, simultaneously performing *Feature Sparsification* and *Amplification*. To characterize this behavior, we introduce two complementary metrics. The first is the ***Spatial Standard Deviation (SSD)***,σ_c, which measures the spatial variability of each feature channel c :

$$\sigma_c = \sqrt{\frac{1}{H \cdot W}\sum_{h=1}^{H}\sum_{w=1}^{W}\left(f_{h,w,c} - \mu_c\right)^2}, \qquad (4)$$

Here, $f_{h,w,c}$ denotes the activation at spatial location (h, w) in channel c, and μ_c is its mean activation across the feature map. Intuitively, channels with high σ_c exhibit strong spatial variation and capture rich geometric structures across spatial regions, whereas very low σ_c indicates redundancy or noise that should be suppressed. In the context of depth estimation, spatially coherent yet non-uniform activations typically correspond to object boundaries, surface discontinuities, and geometric contours, rather than stochastic texture noise.

The second metric, the **Channel Masking Ratio**, is defined as the proportion of channels with σ_c below a threshold $\epsilon = 1.0 \times 10^{-4}$, providing a global measure of feature sparsity. We observe that the overall sparsity trends remain stable across a reasonable range of ϵ values, and thus fix this threshold for all analyses. It is important to note that the Masking Ratio is used solely as a post-hoc diagnostic metric and does not introduce any hard masking during the forward pass.

Together, these two metrics characterize the encoder's role as a channel selector: SSD identifies channels that are amplified to preserve key geometric cues, while the Masking Ratio reflects the extent of redundancy suppression. As summarized in Table I, raw DINOv2 features are dense (Masking Ratio = 0). After fusion, the encoder masks out approximately 77% of channels while simultaneously amplifying the remaining ones, boosting the maximum SSD from 1–5 to over 100. Importantly, this sparsification does not degrade representational capacity, as the preserved channels exhibit substantially

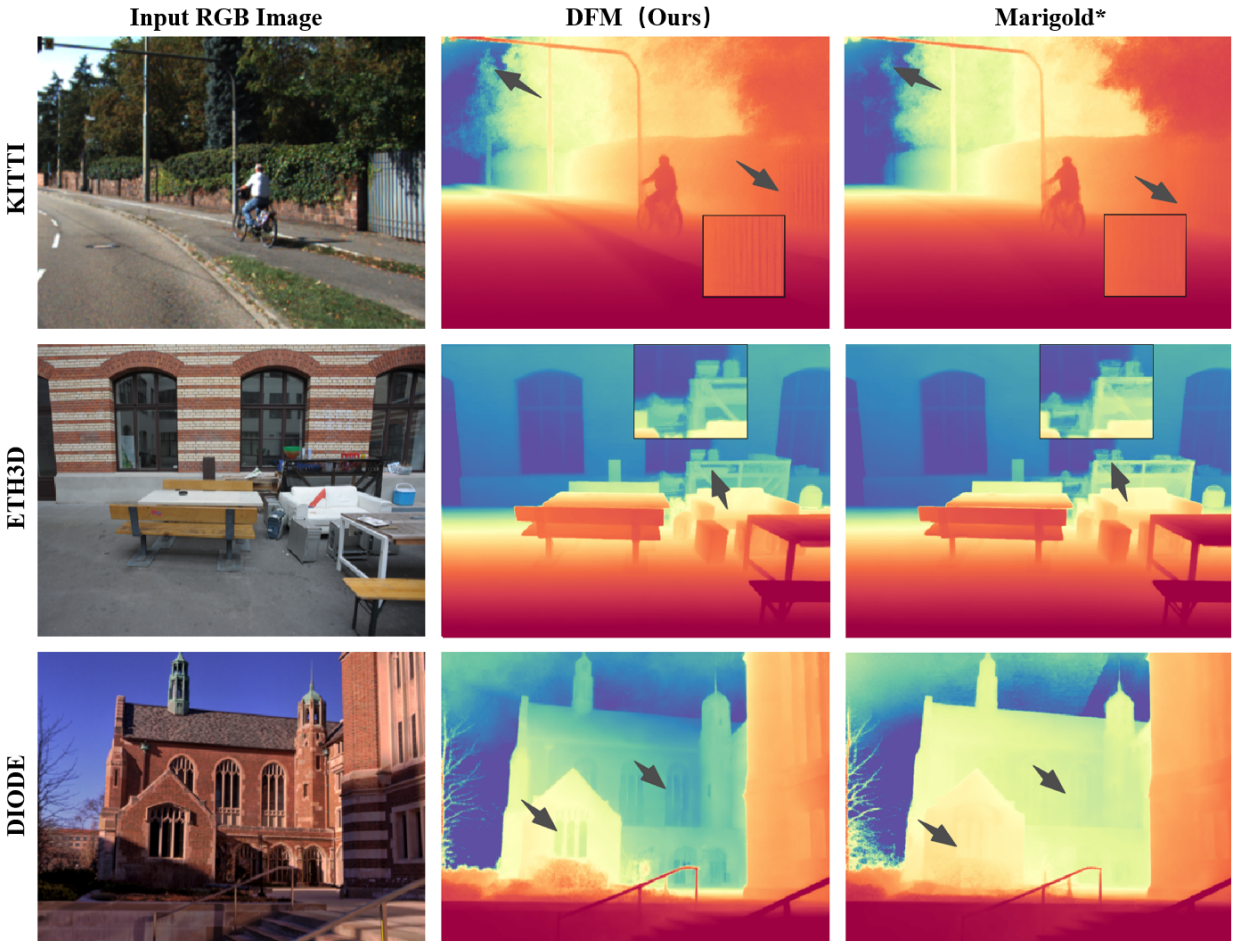

Fig. 3. Qualitative comparison on outdoor benchmarks. Arrows indicate key regions for comparison, and the corresponding local details are shown in enlarged views to highlight structural differences.

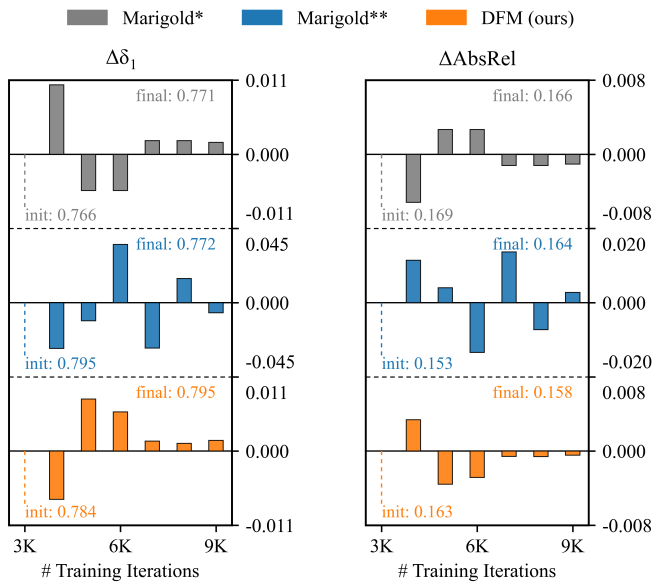

Fig. 4. Step-wise ablation on 100 KITTI validation images (VKITTI training). Left: $\Delta\delta_1$ (+); right: ΔAbsRel (-) over training iterations for Marigold*, Marigold** (+ DINOv2 final-layer feature), and our full DFM, with initial and final scores annotated.

increased spatial expressiveness. This dual process effectively filters noise and magnifies salient geometric structures, yielding a discriminative geometry-aware representation that supports robust depth estimation in downstream tasks.

III. Experiments

A. Implementation Details

Our Hierarchical ViT Encoder is built on DINOv2 (ViT-L/14) [22]. For an input image of size 518×518, the patch-level feature map has a resolution of 37×37. We extract features from the 4th, 11th, 17th, and 23rd blocks of DINOv2 (corresponding to the layer indices j in Eq. (1)), and reorganize and upsample these 1024-dimensional patch features to form $z^{(i)}$ at resolutions 148×148, 74×74, 37×37, and 18×18, respectively, with all channels projected to 512 dimensions for architectural consistency. After multi-channel residual fusion, the final fused feature $f^{(0)} \in \mathbb{R}^{512\times148\times148}$ is downsampled and expanded along the channel dimension to obtain $F \in \mathbb{R}^{1024\times37\times37}$,which is then flattened to produce the token sequence $T \in \mathbb{R}^{B\times N\times1024}$, where $N = 37 \times 37$. The sequence T is used as the Key/Value input to the cross-attention layers of Marigold's U-Net. We fine-tune the model with a learning rate of 6.0×10^{-5} for 24K iterations, reaching optimal performance at approximately 22K iterations.

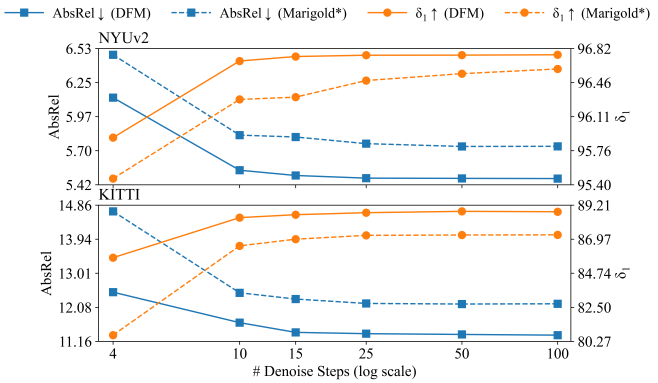

Fig. 5. Performance on NYUv2 and KITTI as a function of the number of denoising steps (ensemble size fixed to 1).

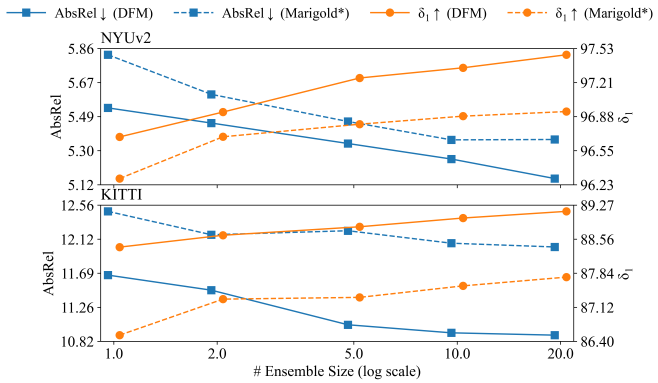

Fig. 6. Performance on NYUv2 and KITTI as a function of the ensemble size (denoising steps fixed to 10). Markers are slightly horizontally offset for visual clarity.

B. Experimental Results

We evaluate DFM and a reproduced Marigold model on indoor (NYUv2 [23], ScanNet [24]) and outdoor (KITTI [25], ETH3D [26], DIODE [27]) benchmarks. Following the standard evaluation protocol, each model is fine-tuned on the official training split of a dataset and tested on the corresponding non-overlapping test split, ensuring fairness and reproducibility across datasets and experimental settings.

Quantitative Comparison. Table II shows that DFM consistently outperforms the Marigold baseline on outdoor benchmarks. On KITTI, DFM raises δ_1 accuracy from 90.90% to 92.15% and reduces AbsRel from 10.03% to 9.82%, with similar gains on ETH3D and DIODE. These results indicate that incorporating DINOv2 features strengthens geometric reasoning in open environments effectively.

For indoor datasets, the trends are more nuanced. On NYUv2, DFM achieves a noticeable improvement over Marigold, whereas on the more diverse ScanNet benchmark it shows a slight performance drop. We attribute this behavior to differences in RGB image quality: compared to relatively clean datasets such as NYUv2, ScanNet is acquired via RGB-D scanning and often contains motion blur, weak textures, and reconstruction artifacts, which can degrade the quality and consistency of DINOv2-derived geometric tokens. While cross-attention effectively integrates high-quality semantic and geometric cues in well-textured scenes, degraded RGB inputs may propagate inconsistencies. Consequently, under these conditions, the fused representation does not exhibit a stable and consistent advantage over the baseline.

Qualitative Comparison. Fig. 3 demonstrates the qualitative advantages of DFM on challenging outdoor scenes, with improved texture fidelity and contour completeness. In KITTI (row 1), DFM recovers fine-grained details such as fence slats and tree textures, while Marigold produces noticeably blurred structures. For contour preservation, DFM maintains sharper boundaries: it separates furniture from the background in ETH3D (row 2) and accurately captures the recessed geometry of church windows in DIODE (row 3).

C. Ablation Study

Layer selection and layer-fusion strategies. We first examine the case of selecting a single-layer DINOv2 feature and observe markedly unstable training behavior on the KITTI dataset (100 samples), validating the critical role of our Hierarchical ViT Encoder in practice. Specifically, we compare a variant that directly feeds raw DINOv2 final-layer features into the model against the full DFM. Fig. 4 uses a waterfall plot to visualize how performance evolves during training, where the initial and final scores are annotated, and each bar shows the change in δ_1 and AbsRel between successive checkpoints. Although directly incorporating raw features yields initial gains, these improvements are quickly offset by substantial negative fluctuations, causing δ_1 and AbsRel to exhibit pronounced instability rather than stable, sustained improvement.

This instability is consistently observed across shallow, intermediate, and high-level DINOv2 features. Although these single-layer variants show only minor differences in test performance, they share similarly unstable optimization behavior. We further compare three fusion strategies: single-layer fusion, multi-layer linear summation, and the proposed Hierarchical ViT Encoder. As shown in Table I, DINOv2 features contain rich structural cues but also substantial noise, making direct integration into Marigold challenging. As a result, directly summing multi-layer features often destabilizes training, with many runs suffering gradient explosion between 2,000 and 8,000 iterations; even single-layer fusion variants are prone to divergence. In contrast, the full DFM model with the Hierarchical ViT Encoder remains stable throughout training and achieves the best final performance ($\delta_1 \approx 0.795$).

Denoising steps. We conduct an ablation study on the number of denoising steps at test time and compare DFM with the Marigold baseline on NYUv2 and KITTI, with the ensemble size fixed to 1 for computational efficiency. Fig. 5 reports AbsRel and δ_1 as functions of the number of denoising steps; all curves are computed on randomly sampled subsets of the test sets rather than the full benchmarks. Increasing the number of denoising steps consistently improves performance for both methods, yielding lower AbsRel and higher δ_1, with the most pronounced gains observed up to 50 steps and only

marginal improvements beyond this point on both datasets. Across all evaluated settings, DFM consistently achieves lower AbsRel and higher δ_1 than Marigold. Notably, this advantage remains evident even with a relatively small number of steps (e.g., 4–10), indicating that the proposed fusion mechanism enables more effective utilization of each denoising iteration and remains beneficial under tight computational budgets.

Ensemble size. We further examine the effect of ensemble size on performance and assess whether the gains of DFM persist under different ensembling settings. To this end, we fix the number of denoising steps to 10 and vary the ensemble size. As shown in Fig. 6, increasing the ensemble size generally improves accuracy for both methods, but exhibits clear diminishing returns once the ensemble size exceeds 10. Similar to Fig. 5, all curves are computed on randomly sampled subsets of the NYUv2 and KITTI test sets rather than the full benchmarks. DFM maintains consistent advantages over Marigold across all ensemble sizes on both NYUv2 and KITTI, indicating that the observed gains are robust to typical test-time ensembling choices rather than being tied to a specific configuration.

IV. Conclusion

In this work, we propose DFM, a monocular depth estimation framework that integrates geometry-aware features from DINOv2 into the Stable-Diffusion-based Marigold pipeline through a Hierarchical ViT Encoder. Unlike naive feature fusion, the proposed encoder selectively enhances informative channels while suppressing redundant ones, producing a more discriminative representation sensitive to geometric structure. This design improves object boundary delineation and surface texture recovery, achieving lower AbsRel and higher δ_1 across multiple outdoor benchmarks. Extensive ablation studies further show that both the hierarchical encoder and the geometric features are essential for stable training and strong depth estimation performance. Future work will explore more efficient inference strategies and stronger geometric priors to further improve overall performance and robustness.

References

[1] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, Canada, 2014, pp. 2366–2374.

[2] X. Xu, Z. Chen, and F. Yin, "Monocular depth estimation with multi-scale feature fusion," *IEEE Sig. Process. Lett.*, vol. 28, pp. 678–682, 2021.

[3] K. Li, Z. Fu, H. Wang, Z. Chen, and Y. Guo, "Adv-depth: Self-supervised monocular depth estimation with an adversarial loss," *IEEE Sig. Process. Lett.*, vol. 28, pp. 638–642, 2021.

[4] R. Marsal, F. Chabot, A. Loesch, W. Grolleau, and H. Sahbi, "Monoprob: Self-supervised monocular depth estimation with interpretable uncertainty," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, 2024, pp. 3637–3646.

[5] X. Hu, C. Zhang, Y. Zhang, B. Hai, K. Yu, and Z. He, "Learning to adapt CLIP for few-shot monocular depth estimation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Waikoloa, HI, USA, 2024, pp. 5594–5603.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[7] Z. Li, Z. Chen, X. Liu, and J. Jiang, "DepthFormer: Exploiting long-range correlation and local information for accurate monocular depth estimation," *Mach. Intell. Res.*, vol. 20, no. 6, pp. 837–854, Dec. 2023.

[8] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6540–6551.

[9] B. Ke, A. Obukhov, S. Huang, N. Metzger, R. C. Daudt, and K. Schindler, "Repurposing diffusion-based image generators for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2024, pp. 9492–9502.

[10] S. Patni, A. Agarwal, and C. Arora, "EcoDepth: Effective conditioning of diffusion models for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, 2024, pp. 28 285–28 295.

[11] D. Zavadski, D. Kalšan, and C. Rother, "PrimeDepth: Efficient monocular depth estimation with a stable diffusion preimage," in *Proc. Asian Conf. Comput. Vis.*, Hanoi, Vietnam, 2024, pp. 922–940.

[12] J. Wang, C. Lin, L. Nie, K. Liao, S. Shao, and Y. Zhao, "Digging into contrastive learning for robust depth estimation with diffusion models," in *Proc. 32nd ACM Int. Conf. Multimedia*, Melbourne, VIC, Australia, 2024, pp. 4129–4137.

[13] Y. Duan, X. Guo, and Z. Zhu, "DiffusionDepth: Diffusion denoising approach for monocular depth estimation," in *Proc. Eur. Conf. Comput. Vis.*, Milan, Italy, 2024, pp. 432–449.

[14] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 10 684–10 695.

[15] J. Lee, M. Heo, K. Kim, and C. Kim, "BTS: A scale-aware network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 5435–5444.

[16] H. Sun, S. Li, J. Chen, Z. Liu, and J. Yu, "High-resolution depth estimation via deep propagation and multi-scale fusion," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 3, 2021, pp. 2663–2671.

[17] M. Bhat, A. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4009–4018.

[18] A. Li, X. Xie, C. Zhang, C. Dong, Q. Dai, and F. Fang, "BinsFormer: Revisiting adaptive bins for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, New Orleans, LA, USA, 2022, pp. 6979–6989.

[19] J. Zhang, C. Herrmann, J. Hur, L. Polania, V. Jampani, D. Sun, and M.-H. Yang, "A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence," in *Proc. Adv. Neural Inf. Process. Syst.*, New Orleans, LA, USA, 2023, pp. 45 533–45 547.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Munich, Germany, 2015, pp. 234–241.

[21] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 2117–2125.

[22] M. Oquab *et al.*, "DINOv2: Learning robust visual features without supervision," arXiv preprint arXiv:2304.07193, 2023. [Online]. Available: https://arxiv.org/abs/2304.07193

[23] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, Florence, Italy, 2012, pp. 746–760.

[24] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "ScanNet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 5826–5836.

[25] A. Geiger, P. Lenz, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.

[26] T. Schöps, J. L. Schönberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, 2017, pp. 3240–3249.

[27] I. Vasiljevic *et al.*, "DIODE: A dense indoor and outdoor depth dataset," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, Aspen, CO, USA, 2020, pp. 110–119.