

ApDepth V2: Geometry-Aware Refinement for Single-Step Monocular Depth Estimation

JIawei WANG^{✉*}†, Shenyang Jianzhu University, China

MINGBO LEI*, Shenyang Jianzhu University, China

JINGXUAN WANG, Beijing University of Chinese Medicine, China

YIBO CHEN, Zhejiang Sci-Tech University, China

LIANG YAO[‡] and YUSU LIANG[‡], Shenyang Jianzhu University, China

SHUAI YUAN[§], Shenyang Jianzhu University, China

ApDepth combines a pretrained depth prior with a diffusion U-Net for deterministic single-step monocular depth estimation. Its original recipe uses a separate feature-alignment stage followed by reconstruction and late frequency-domain refinement. This update revisits local geometry supervision while retaining the single-step predictor. We introduce a continuous fine-tuning recipe initialized directly from Stable Diffusion 2.1, with persistent reconstruction constraints, delayed decoder adaptation, and Spatially Debaised Window Transport (SDWT). Building on the local matching formulation of Marigold V2, SDWT combines depth and spatial costs, subtracts self-transport terms, evaluates two offset window grids, and weights windows by valid support. These choices are intended to tolerate small supervision misalignments without discarding local geometry. We also correct the auxiliary-prior interface and separate optional measured far-depth supervision from the default objective. This working manuscript specifies the implemented method and a controlled evaluation protocol; benchmark improvements and efficiency gains remain to be established experimentally.

Additional Key Words and Phrases: Monocular depth estimation, single-step prediction, local optimal transport, decoder adaptation

1 Introduction

Pretrained image generators provide a useful starting point for monocular depth estimation, as demonstrated by Marigold and its extensions [18, 19]. Our earlier ApDepth [38] combines this generative initialization with a frozen Depth Anything V2 prior [44]. A deterministic U-Net predicts a depth latent in one pass, while two-stage training first aligns representations and then emphasizes depth reconstruction and frequency content. The architecture avoids iterative denoising, but the recipe leaves several questions open: is a separate alignment stage necessary, can spectral refinement reliably localize edges, and does a frozen image decoder limit depth-specific adaptation?

The present update focuses on these questions. A Fourier-magnitude objective constrains spatial frequency content but does not uniquely determine its location: even a translated signal can preserve its magnitude spectrum. Conversely, pointwise losses penalize a small displacement of an otherwise sharp depth boundary. Such conflicts matter around thin structures and ambiguous annotations. Marigold V2 [23] addresses this issue using local Sinkhorn matching, which relaxes exact pixel

*Both authors contributed equally to this research.

†Also with InfiniFlow (Shanghai) Information Technology Co., Ltd..

‡These authors inspired our work and provided valuable suggestions.

§The author would like to acknowledge Shuai Yuan for providing experimental equipment and partial guidance for this work.

correspondence. We adopt this motivation and examine how to retain spatial information in the relaxed objective.

ApDepth V2 keeps the prior-conditioned U-Net and replaces mandatory two-stage initialization with one continuous SD2.1 fine-tuning run. Reconstruction and gradient supervision establish the initial mapping; SDWT and decoder adaptation are introduced later. SDWT extends local depth matching with a displacement penalty, a standard self-cost correction, offset windows, and valid-support aggregation. Frequency supervision remains an optional auxiliary term, rather than becoming the only objective.

This scope differs from our companion ApDepth-G manuscript [39], which studies sky-depth collapse using validity-guided completion in a multi-step diffusion process. Here, the main objective supervises *observed local geometry*. An optional far-depth branch supplies targets where the dataset identifies known far values; it is not a semantic sky detector. Our contributions are the implemented refinement formulation, its continuous training recipe, and an evaluation design separating geometry supervision from interface corrections. Their empirical value remains open in this pre-experiment draft.

2 Related Work

Depth predictors and generalization. Supervised depth prediction has evolved from multiscale regression and ordinal or planar representations [6, 8, 21, 47] to mixed-dataset training and transformer-based models [4, 5, 27, 28]. Self-supervised approaches exploit image reconstruction [13, 14]. Depth Anything and Depth Anything V2 [43, 44] use large-scale visual representations, including DINOv2 [22], to support broad generalization. Metric reconstruction and fine-boundary prediction motivate complementary designs such as UniDepth and Depth Pro [1, 26]. We use a pretrained relative-depth predictor as a condition, not as metric ground truth.

Repurposing generative models. Latent diffusion [30] and U-Net backbones [31] underpin Marigold, GeoWizard, and DepthFM [9, 15, 18]. Efficient adaptations include E2E fine-tuning, GenPercept, and Lotus [11, 17, 42]; BetterDepth [48] conditions refinement on an existing depth estimate. Diffusion transformers [24], image-editing models [40], and recent depth systems such as Lotus-2, Pixel-Perfect Depth, and InfiniDepth [16, 41, 45] broaden the accuracy–detail trade-off. We retain SD2.1 to isolate supervision changes from backbone capacity.

Representations, boundaries, and transport. REPA and iREPA [34, 46] motivate representation alignment; DepthMaster [36] and the original ApDepth apply related ideas to depth. SharpDepth [25] and boundary-oriented evaluation [20] emphasize quality beyond aggregate error. Our closest methodological reference is Marigold V2’s SinkLoss [23], based on local entropy-regularized matching [2, 35]. We extend its application with spatial costs and masked aggregation, and adopt the established Sinkhorn-divergence self-cost correction [7]. Neither local transport nor debiasing alone is claimed as a new general principle.

3 Method

3.1 Single-Step Predictor and Prior Interface

For RGB input I , let $P(I)$ be the frozen DA2-Giant prediction, $\mathcal{E}$ the frozen VAE encoder, and $\mathcal{D}_\phi$ the decoder. With R denoting replication into three channels, the predictor is

$$z_I = \mathcal{E}(I), \quad z_p = \mathcal{E}(R(p)), \quad \hat{z} = f_\theta([z_p, z_I], 1, e_0), \quad \hat{d} = \mathcal{D}_\phi(\hat{z}). \quad (1)$$

The VAE encoder uses its posterior mean with latent scale 0.18215; decoder outputs are averaged over three channels. The U-Net accepts eight channels, ordered as prior then RGB, and directly predicts a depth latent at fixed timestep 1. No random depth initialization or scheduler integration

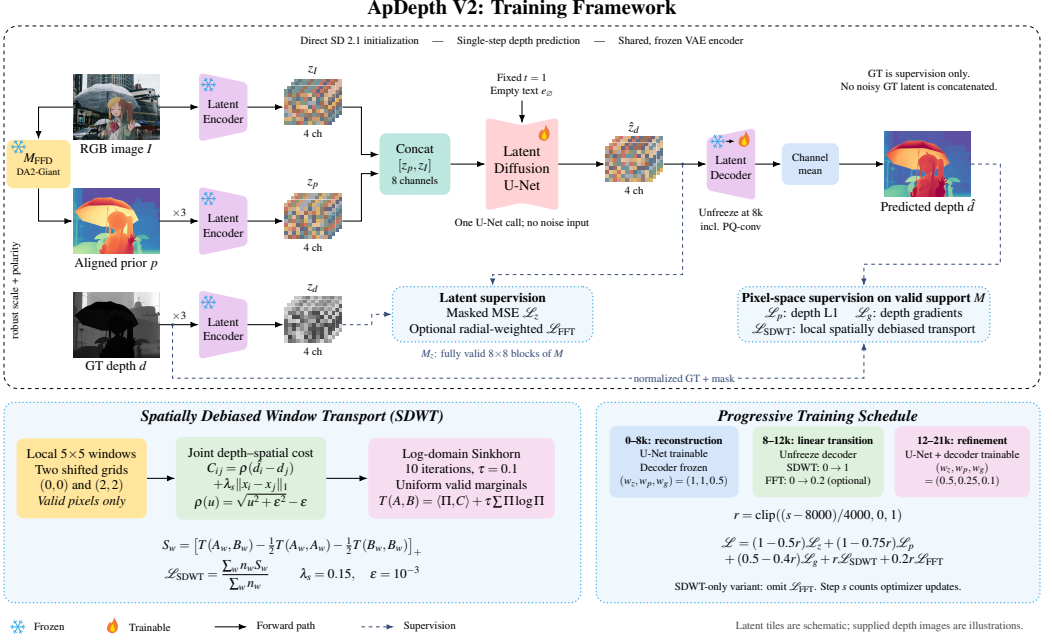


Fig. 1. Training framework for ApDepth v2

is used. Thus single-step means one U-Net evaluation; DA2 and both encoding paths still contribute to end-to-end cost.

For metric depth D and valid mask M , we retain clipped *linear* depth normalization:

$$d = 2 \cdot \text{clip}_{[0,1]} \left(\frac{D - q_{.02}^D}{q_{.98}^D - q_{.02}^D} \right) - 1, \quad (2)$$

where quantiles use positive valid depths. This representation does not provide calibrated metric scale at inference. Degenerate valid-depth ranges must be excluded or handled explicitly in preprocessing.

The updated DA2 interface converts ApDepth RGB from $[-1, 1]$ to $[0, 1]$ and applies the ImageNet channel normalization expected by its encoder. Its disparity-like output $u = P(I)$, where larger means nearer, is mapped to

$$p = 1 - 2 \cdot \text{clip}_{[0,1]} \left(\frac{u - q_{.02}^u}{\max(q_{.98}^u - q_{.02}^u, 10^{-6})} \right). \quad (3)$$

These quantiles use the per-image prior, not ground truth. This corrects ordering and dynamic range; it is not metric inversion of disparity. Controlled training and evaluation must share this correction: replacing the interface only at test time changes the conditioning distribution of an older checkpoint.

3.2 Spatially Debiased Window Transport

Depth-only local matching can assign the same cost to different spatial permutations inside a window. This permits annotation tolerance but weakens localization. We retain the relaxation while making displacement costly.

For a $K \times K$ window w , remove invalid positions before constructing transport matrices. Let $A_w = \{(\hat{d}_i, x_i)\}_{i=1}^{n_w}$ and $B_w = \{(d_i, x_i)\}_{i=1}^{n_w}$ contain prediction and target samples at shared valid positions. Coordinates are local pixel indices divided by $2(K-1)$, so their maximum pairwise Manhattan distance is one. The joint cost is

$$C_{ij} = \rho_\epsilon(\hat{d}_i - d_j) + \lambda_s \|x_i - x_j\|_1, \quad \rho_\epsilon(a) = \sqrt{a^2 + \epsilon^2} - \epsilon. \quad (4)$$

The smooth absolute-depth penalty reduces sensitivity to large residuals compared with squared error. The spatial term discourages distant exchanges, without imposing a hard displacement threshold.

With uniform marginals $\mathbf{1}/n_w$, define

$$\mathcal{T}_\tau(A, B) = \min_{\Pi \geq 0: \Pi \mathbf{1} = \Pi^\top \mathbf{1} = \mathbf{1}/n_w} \left[\langle \Pi, C \rangle + \tau \sum_{ij} \Pi_{ij} \log \Pi_{ij} \right]. \quad (5)$$

We approximate this objective using ten log-domain Sinkhorn normalization iterations, differentiating through the iterations. Following Sinkhorn divergences [7], the window loss is

$$S_w = \left[\mathcal{T}_\tau(A_w, B_w) - \frac{1}{2} \mathcal{T}_\tau(A_w, A_w) - \frac{1}{2} \mathcal{T}_\tau(B_w, B_w) \right]_+. \quad (6)$$

Self-terms remove the entropic self-cost; identical inputs give zero up to numerical precision. The implementation clamps finite-iteration negative values to zero. A fixed iteration count does not establish exact metric guarantees or convergence. The target self-term is evaluated without a gradient graph.

We evaluate grids with offsets $(0, 0)$ and $(\lfloor K/2 \rfloor, \lfloor K/2 \rfloor)$. Padded positions are invalid. For nonempty windows $\mathcal{W}$ from both grids,

$$\mathcal{L}_{\text{SDWT}} = \frac{\sum_{w \in \mathcal{W}} n_w S_w}{\sum_{w \in \mathcal{W}} n_w}. \quad (7)$$

The second grid allows some boundaries split by the first grid to be compared within a window; it reduces dependence on one partition without ensuring translation invariance. Support weighting prevents a one-pixel window from receiving the same weight as a fully observed window. Empty support contributes zero. Defaults are $K = 5$, $\tau = 0.1$, $\lambda_s = 0.15$, and $\epsilon = 10^{-3}$. Windows are grouped by support size and processed in chunks of 512 with activation checkpointing. Debiasing and offset windows add training computation but no inference module.

3.3 Continuous Training and Decoder Adaptation

We initialize from base SD2.1 without the original feature-alignment checkpoint. The four-channel input convolution is duplicated and halved to accept eight channels. The text encoder, DA2, VAE encoder, and quantization convolution remain frozen. Initially only the U-Net is trained; the VAE decoder and post-quantization convolution become trainable after 8,000 completed optimizer updates.

Let $\mathcal{L}_z$ be masked latent MSE and $\mathcal{L}_p$ masked pixel L_1 . A latent cell is supervised only if every pixel in its associated 8×8 mask block is valid; this rule does not confine the VAE receptive field to that block. Gradient supervision uses first differences with both endpoints valid:

$$\mathcal{L}_g = \frac{\sum_{(i,j) \in \mathcal{N}_M} |(\hat{d}_i - \hat{d}_j) - (d_i - d_j)|}{\max(|\mathcal{N}_M|, 1)}, \quad (8)$$

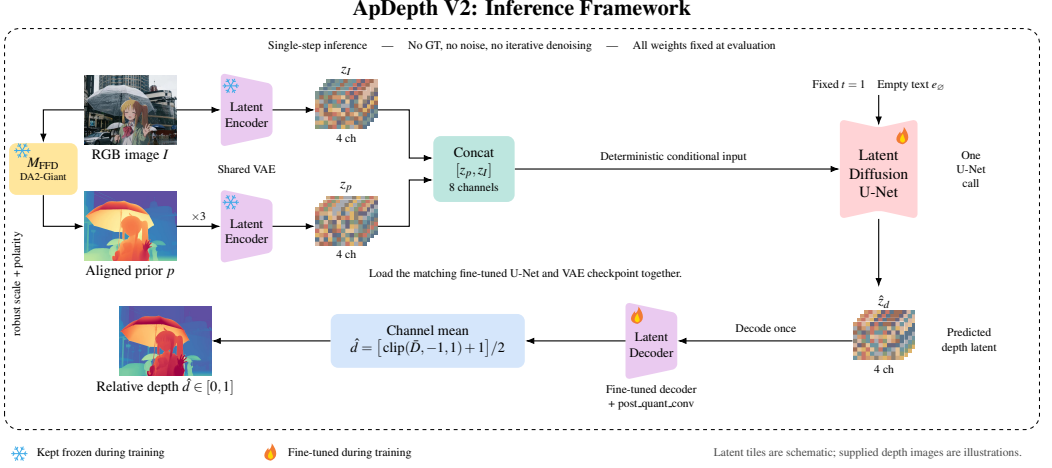


Fig. 2. ApDepth v2 inference framework

where N_M contains horizontal and vertical adjacent valid pairs. For completed update count s , let $r(s) = \text{clip}_{[0,1]}((s - 8000)/4000)$. The implemented objective is

$$\begin{aligned} \mathcal{L}(s) = & (1 - 0.5r)\mathcal{L}_z + (1 - 0.75r)\mathcal{L}_p + (0.5 - 0.4r)\mathcal{L}_g \\ & + r\mathcal{L}_{\text{SDWT}} + \lambda_F(s)\mathcal{L}_F. \end{aligned} \quad (9)$$

The basic configuration sets $\lambda_F = 0$; the FFT variant uses $\lambda_F = 0.2r$. Here $\mathcal{L}_F$ is the inherited mean absolute difference of centered latent Fourier magnitudes, with radial high-frequency weighting. Invalid latent locations are zeroed before the FFT. The transform is unnormalized and its scale depends on resolution and mask structure; a small coefficient alone does not ensure a small gradient contribution. Loss and gradient scales should therefore be monitored.

Equation (9) retains reconstruction throughout refinement, replacing the original late switch to frequency-only supervision. Decoder adaptation lets pixel-space objectives change the latent-to-depth mapping while the target encoder stays fixed. It may reduce reconstruction artifacts but cannot restore information already discarded by the latent representation. The learned U-Net and VAE must be loaded together; the updated checkpoint loader supports this pairing.

3.4 Optional Known-Far Supervision

A separate post-training recipe, inspired by ApDepth-G [39], uses native Virtual KITTI depths in $[80, 655.35]$ m, excluded from the standard 80 m valid range, as known-far support F . These positions receive target +1 before VAE encoding. Quantiles still use ordinary valid depths. Additional latent supervision covers cells fully supported by $M \cup F$ but not fully supported by M , including mixed valid/far cells. A separately normalized latent-MSE plus pixel- L_1 term is weighted by 0.5.

This branch is disabled in both default SDWT configurations. Unknown invalid pixels remain unsupervised, and known-far values may describe buildings as well as sky. The branch provides a bounded far-ordering cue, not evidence that SDWT resolves sky collapse. Its effect requires a separate ablation. The legacy post-training configuration also needs explicit adaptation before combination with a learned SDWT decoder.

Table 1. Planned zero-shot comparison using the original grouped table style. All values await matched-protocol evaluation; — means not measured. Columns report 100 AbsRel and 100 δ_1 . DIODE requires native validity, not the legacy edge-filtered mask.

Method	NYUv2		KITTI		ETH3D		ScanNet		DIODE	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
<i>Discriminative reference</i>										
DA2 [44]	—	—	—	—	—	—	—	—	—	—
<i>Generative-pretrained references</i>										
Marigold [18]	—	—	—	—	—	—	—	—	—	—
DepthMaster [36]	—	—	—	—	—	—	—	—	—	—
Marigold V2 [23]	—	—	—	—	—	—	—	—	—	—
ApDepth [38]	—	—	—	—	—	—	—	—	—	—
ApDepth-G [39]	—	—	—	—	—	—	—	—	—	—
ApDepth V2	—	—	—	—	—	—	—	—	—	—
ApDepth V2 + FFT	—	—	—	—	—	—	—	—	—	—

4 Experimental

Status and training settings. All table entries marked — are pending measurements; no new result is reported. The recipe specifies 21,000 updates, effective batch size 32, Adam, initial learning rates 3×10^{-5} for the U-Net and 3×10^{-6} for the decoder, 100 warm-up steps, exponential decay, and gradient clipping at 1.0. Training samples Hypersim [29] and Virtual KITTI [10] in a 9:1 ratio using the original manifests. Hypersim is resized to 480×640 ; Virtual KITTI uses the inherited benchmark crop. Hardware, runtime, and retained sample counts will be recorded with the runs. These settings are starting points, not validated optima.

Accuracy and boundaries. We plan zero-shot evaluation on NYUv2 [33], KITTI [12], ETH3D [32], ScanNet [3], and DIODE [37]. Each prediction is aligned to valid ground-truth depth with least-squares scale and shift, then clipped to the dataset range. Table 1 retains the original grouped dataset columns and paired AbsRel/ δ_1 metrics, expressed as percentages. Splits, crops, alignment, resolution, and prior checkpoints must match across methods.

The inherited DIODE loader additionally excludes pixels whose ground-truth Sobel-gradient magnitude exceeds 0.3. This legacy edge-filtered protocol differs from the standard validity mask. Our primary comparison requires native-validity evaluation; historical edge-filtered scores must be reported separately. This evaluator change remains to be implemented before experiments. Boundary-metric utilities also exist but are not called by the default evaluator. Boundary F1 should use valid adjacent pairs, fixed thresholds, and consistent alignment, accompanied by thin-structure crops. Aggregate depth error alone cannot establish improved localization.

Controlled ablations. Table 2 separates historical replication from method ablations. A historical checkpoint and interface document the original system; a retrained original recipe with the corrected interface isolates the implementation correction. Component ablations then share direct SD2.1 initialization, data order, update budget, and corrected DA2 interface. The supplied SDWT ablation configurations inherit the *no-FFT* base; the FFT variant is tested separately. The curriculum remains fixed unless the schedule itself is the experimental factor.

The supplied no-debias switch returns the unregularized cost of the entropy-fitted cross plan. It changes both self-cost subtraction and the scalar objective, so it is not a pure self-term ablation. A

Table 2. Planned ablations, with pending accuracy values in the same percentage convention as Table 1. SDWT uses no FFT. Transport controls marked † require additional implementation.

Setting	NYUv2		KITTI	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
Original recipe, historical interface	—	—	—	—
Original recipe, corrected interface	—	—	—	—
SDWT (full, no FFT)	—	—	—	—
Without transport	—	—	—	—
Without spatial cost	—	—	—	—
Cross transport cost only	—	—	—	—
Regularized cross objective only †	—	—	—	—
Single grid	—	—	—	—
Equal window weights	—	—	—	—
Frozen decoder	—	—	—	—
SinkLoss reference †	—	—	—	—
SDWT + FFT	—	—	—	—

Table 3. Planned cost accounting at matched precision, resolution, batch size, and hardware. All values are pending. End-to-end latency includes DA2 and VAE processing.

Setting	Training GPU-hours	Peak training memory (GB)	U-Net latency (ms)	End-to-end latency (ms)
Original ApDepth	—	—	—	—
ApDepth V2	—	—	—	—
ApDepth V2 + FFT	—	—	—	—

control retaining the regularized cross objective is needed to isolate debiasing. A faithful SinkLoss reference should separately reproduce Marigold V2’s cost, masking, grid, and iteration settings. Main comparisons should use multiple training seeds and report per-seed results and uncertainty.

Efficiency and targeted diagnostics. Table 3 distinguishes training simplification from inference speed. Time the prior, encoders, U-Net, decoder, and total pipeline on identical hardware and resolution, with warm-up and device synchronization. Report training GPU-hours, throughput, and peak allocated memory, including the original feature-alignment stage. SDWT adds no inference-time loss computation, but its three transport solves per window and two grids increase training cost. Removing an initialization stage alone does not establish lower total GPU-hours.

Targeted diagnostics should test small boundary shifts, larger within-window permutations, window-grid offsets, and different valid-support fractions. These perturbations test intended loss behavior without substituting for trained-model evaluation. Optional far supervision requires separate known-far and sky diagnostics alongside foreground accuracy, since sensor masks often exclude the regions at issue.

5 Limitations and Conclusion

ApDepth V2 updates the training of a prior-conditioned single-step depth predictor through geometry-aware local transport, persistent reconstruction, and delayed decoder adaptation. The

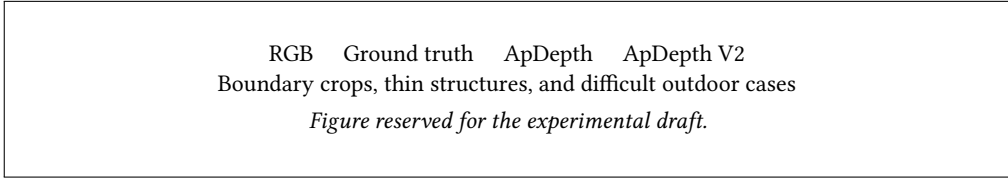


Fig. 3. Reserved qualitative comparison. Use identical crops and depth color normalization, and show invalid ground-truth regions. No predictions are presented in this draft.

implementation motivates tests of boundary localization, annotation tolerance, and decoder artifacts, but does not yet establish benchmark improvements.

The method retains an expensive and imperfect external prior. Local transport permits rearrangements and may suppress genuine tiny structures; spatial weighting trades tolerance against localization. Masked supervision does not constrain wholly unobserved sky, while decoder adaptation may overfit. Finite transport iterations, mask-induced spectral artifacts, and removing explicit representation alignment introduce further trade-offs. The planned experiments will determine whether the combination offers a useful accuracy–detail–cost balance.

References

- [1] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. 2025. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. In *ICLR*. <https://arxiv.org/abs/2410.02073>
- [2] Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems* 26 (2013).
- [3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2432–2443.
- [4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.
- [5] Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. 2021. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 10786–10796.
- [6] David Eigen, Christian Puhres, and Rob Fergus. 2014. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems* 27 (2014).
- [7] Jean Feydy, Thibault Sjourne, Franois-Xavier Vialard, Shun ichi Amari, Alain Trounev, and Gabriel Peyr. 2019. Interpolating between Optimal Transport and MMD using Sinkhorn Divergences. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 89)*. PMLR, 2681–2690. <https://proceedings.mlr.press/v89/feidy19a.html>
- [8] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. 2018. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2002–2011.
- [9] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. 2024. GeoWizard: Unleashing the Diffusion Priors for 3D Geometry Estimation from a Single Image. In *European Conference on Computer Vision*. 241–258.
- [10] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. 2016. Virtual Worlds as Proxy for Multi-Object Tracking Analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [11] Gonzalo Martin Garcia, Karim Knaebel, Christian Schmidt, Daan de Geus, Alexander Hermans, and Bastian Leibe. 2025. Fine-Tuning Image-Conditional Diffusion Models is Easier than You Think. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
- [12] Andreas Geiger, Philip Lenz, and Raquel Urtasun. 2012. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3354–3361.

- [13] Clément Godard, Oisín Mac Aodha, and Gabriel J. Brostow. 2017. Unsupervised Monocular Depth Estimation with Left-Right Consistency. In *CVPR*.
- [14] Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J. Brostow. 2019. Digging into Self-Supervised Monocular Depth Prediction. In *The International Conference on Computer Vision (ICCV)*.
- [15] Ming Gui, Johannes Schusterbauer, Ulrich Prestel, Pingchuan Ma, Dmytro Kotovenko, Olga Grebenkova, Stefan Andreas Baumann, Vincent Tao Hu, and Björn Ommer. 2025. DepthFM: Fast Generative Monocular Depth Estimation with Flow Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 3203–3211.
- [16] Jing He, Haodong Li, Mingzhi Sheng, and Ying-Cong Chen. 2025. Lotus-2: Advancing Geometric Dense Prediction with Powerful Image Generative Model. *arXiv preprint arXiv:2512.01030* (2025).
- [17] Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and YingCong Chen. 2025. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *International Conference on Learning Representations*, Vol. 2025. 89454–89467.
- [18] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2024. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9492–9502.
- [19] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. 2025. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2025).
- [20] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. 2018. Evaluation of CNN-based Single-Image Depth Estimation Methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.
- [21] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. 2019. From Big to Small: Multi-Scale Local Planar Guidance for Monocular Depth Estimation. *arXiv:1907.10326*. <https://arxiv.org/abs/1907.10326>
- [22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, et al. 2024. DINOv2: Learning Robust Visual Features without Supervision. *Transactions on Machine Learning Research* (2024).
- [23] Igor Pavlovic, Thimo Wandel, Anton Obukhov, Luca Bartolomei, Andrey Davydov, Fabio Tosi, Matteo Poggi, Sabine Süsstrunk, and Dengxin Dai. 2026. Marigold V2: Revisiting Diffusion Transformers for Monocular Depth Estimation. *arXiv:2609.08084*. <https://arxiv.org/abs/2609.08084>
- [24] William Peebles and Saining Xie. 2023. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4195–4205.
- [25] Duc-Hai Pham, Tung Do, Phong Nguyen, Binh-Son Hua, Khoi Nguyen, and Rang Nguyen. 2025. Sharpdepth: Sharpening metric depth predictions using diffusion distillation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 17060–17069.
- [26] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. 2024. UniDepth: Universal Monocular Metric Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA.
- [27] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 12179–12188.
- [28] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2020. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence* 44, 3 (2020), 1623–1637.
- [29] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. 2021. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
- [30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. 234–241.
- [32] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. 2017. 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*. 3260–3269.
- [33] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *European Conference on Computer Vision*. Springer, 746–760.
- [34] Jaskirat Singh, Xingjian Leng, Zongze Wu, Liang Zheng, Richard Zhang, Eli Shechtman, and Saining Xie. 2026. What matters for Representation Alignment: Global Information or Spatial Structure?. In *The Fourteenth International Conference on Learning Representations*. <https://openreview.net/forum?id=y0UxFtXqXf>

- [35] Richard Sinkhorn and Paul Knopp. 1967. Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* 21, 2 (1967), 343–348.
- [36] Ziyang Song, Zerong Wang, Bo Li, Hao Zhang, Ruijie Zhu, Li Liu, Peng-Tao Jiang, and Tianzhu Zhang. 2026. Depth-master: Taming diffusion models for monocular depth estimation. *IEEE Transactions on Circuits and Systems for Video Technology* (2026).
- [37] Igor Vasiljevic, Nicholas I. Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. 2019. DIODE: A Dense Indoor and Outdoor DEpth Dataset. *arXiv preprint arXiv:1908.00463* (2019).
- [38] Jiawei Wang, Mingbo Lei, Haoze Shou, Yusu Liang, and Yibo Chen. 2026. ApDepth: Aiming for Precise Monocular Depth Estimation Based on Diffusion Models. (2026). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=7604741 Preprint. Temporary bibliographic record pending SSRN metadata confirmation.
- [39] Jiawei Wang, Mingbo Lei, Liang Yao, Jiayi Li, and Jingxuan Wang. 2026. ApDepth-G: Addressing Sky-Depth Collapse with Validity-Guided Latent Diffusion. (2026). Manuscript in preparation. Temporary reference; arXiv record pending.
- [40] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).
- [41] Gangwei Xu, Haotong Lin, Hongcheng Luo, Xianqi Wang, Jingfeng Yao, Lianghui Zhu, Yuechuan Pu, Cheng Chi, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Sida Peng, and Xin Yang. 2025. Pixel-perfect depth with semantics-prompted diffusion transformers. *Advances in Neural Information Processing Systems* 38 (2025), 174731–174755.
- [42] Guangkai Xu, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, Chunhua Shen, et al. 2025. What matters when repurposing diffusion models for general dense perception tasks?. In *International Conference on Learning Representations*, Vol. 2025. 6786–6799.
- [43] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10371–10381.
- [44] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. 2024. Depth anything v2. *Advances in Neural Information Processing Systems* 37 (2024), 21875–21911.
- [45] Hao Yu, Haotong Lin, Jiawei Wang, Jiaxin Li, Yida Wang, Xueyang Zhang, Yue Wang, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. 2026. InfiniDepth: Arbitrary-Resolution and Fine-Grained Depth Estimation with Neural Implicit Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- [46] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. 2025. Representation Alignment for Generation: Training Diffusion Transformers Is Easier Than You Think. In *International Conference on Learning Representations*.
- [47] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. 2022. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3916–3925.
- [48] Xiang Zhang, Bingxin Ke, Hayko Riemenschneider, Nando Metzger, Anton Obukhov, Markus Gross, Konrad Schindler, and Christopher Schroers. 2024. BetterDepth: Plug-and-Play Diffusion Refiner for Zero-Shot Monocular Depth Estimation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. <https://openreview.net/forum?id=35WwZhksh>