Geometric Distillation from Rectified Stereo: Leveraging Epipolar Cues for Monocular Depth

Jung-Hee Kim[1] and Xiaoming Liu[1,2]

[1] Michigan State University, East Lansing, MI 48824, USA
[2] University of North Carolina at Chapel Hill, Chapel Hill, NC 27514, USA
kimjun84@msu.edu, liuxm@cs.unc.edu

Abstract. Monocular depth foundation models have demonstrated remarkable generalization capabilities across diverse environments. However, they continue to struggle with metric depth estimation in diverse environments. This limitation stems from the inherent scale ambiguity of single-view inference, leading to misaligned scale predictions even when the relative geometry is accurate. Conversely, recent multi-view foundation models leverage cross-view cues to learn robust scene-level geometry and consistent scale. Yet, these benefits typically vanish during single-image inference, as the absence of explicit geometric constraints causes performance to degrade. To bridge this gap, we propose a novel framework that transfers the scale-aware geometric priors of multi-view models into monocular depth foundation models. Specifically, we introduce an Epipolar Distillation (EpiDistill), an approach utilizing Rectified Stereo Tokens, which enables the single-view prediction model to retain epipolar attention patterns and maintain geometric consistency without requiring multi-view inputs at inference. Experimental results demonstrate that our method significantly improves zero-shot metric depth estimation, particularly on challenging datasets like ETH3D and DIODE where scale alignment is critical. Furthermore, our approach is model-agnostic, consistently boosting the performance of state-of-the-art ViT-based models, including UniDepthV2 and DepthPro. *Project Link*

Keywords: Monocular Depth Estimation · Multi-view Geometry · Knowledge Distillation

1 Introduction

Monocular depth foundation models [2,3,14,21,26,27,43,44] have demonstrated remarkable generalization capabilities across diverse environments. Despite being trained on massive datasets, these models inherently struggle to perceive the correct depth scale, a fundamental challenge in metric depth estimation. To resolve this scale ambiguity, early methods [15, 46] leveraged camera intrinsics at inference time to disentangle focal length from depth predictions. Recent approaches have relaxed this constraint by directly regressing camera intrinsics alongside depth [26, 27, 37, 38], relying on large-scale datasets to implicitly learn the correlation between the two.

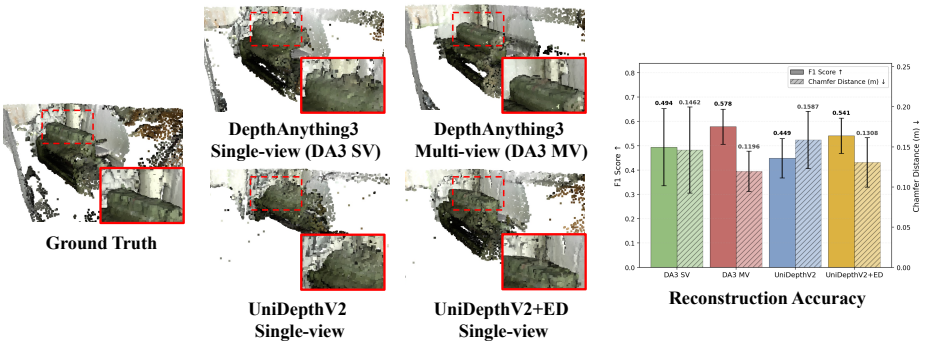

Fig. 1: 3D Scene Reconstruction on ScanNet++ [45]. (Left) Qualitative comparison of point clouds accumulated across multiple sequential frames using single-view (SV) and multi-view (MV) methods. **(Right)** Quantitative evaluation using F1 Score (↑) and Chamfer Distance (↓) on the test set. ED indicates the proposed *EpiDistill* method.

Concurrently, the recent advent of multi-view foundation models [16, 21, 36] has enabled robust, scene-level 3D reconstruction from a sequence of images. These methods exploit task-specific tokens and leverage geometry-aware learning through global cross-attention layers. By processing multi-view inputs in a single feed-forward pass, they successfully establish geometric correspondence and predict a globally consistent metric scale.

However, a critical limitation arises when these multi-view foundation models are tasked with single-view inference. Restricted to single-view inputs, their performance degrades significantly, often falling short of dedicated monocular models. This degradation fundamentally stems from an architectural collapse: in the absence of multi-view input, the cross-view global attention degenerates into a standard intra-frame self-attention. This creates a geometry mismatch between training and inference, leading to the immediate loss of the scale-aware geometric priors learned from multi-view data.

We empirically demonstrate this limitation in Fig. 1. Given multi-view images with camera parameters, a multi-view model such as DepthAnything3 (DA3) [21] explicitly processes 5-frame batches to align the scale, generating well-aligned point clouds in a single feed-forward pass. Conversely, when forced to perform single-view inference and accumulate results sequentially, the model severely struggles with per-frame scale ambiguity, producing misaligned reconstructions. This confirms that vital geometric structures, particularly metric scale, are inherently lost when the multi-view attention mechanism collapses. To date, how to effectively project the scale robustness of multi-view models into single-view prediction remains an open challenge.

To bridge this gap, we introduce *EpiDistill*, a novel geometric distillation framework that explicitly transfers multi-view geometric knowledge to a single-view model. During training, a multi-view model achieves accurate depth per-

ception by utilizing an explicit *Depth-Guided Epipolar Attention*. To process single-view inputs while preserving this learned structure, our single-view model pairs the retained epipolar attention with learnable *Rectified Stereo Tokens*, directly distilling knowledge from the multi-view tokens. These tokens serve as a structural anchor to provide the geometric guidance necessary to effectively leverage the distilled priors. As visually and quantitatively evidenced in Fig. 1, integrating our method with a baseline, UniDepthV2+*EpiDistill* (ED), successfully produces accurate, globally scale-aligned reconstructions from single-view inputs—improving the F1 score by 20.5% and reducing the Chamfer distance by 17.6%. By inheriting the scale robustness inherent to multi-view geometry, our approach significantly elevates the zero-shot metric depth estimation capabilities of state-of-the-art (SoTA) Vision Transformer (ViT) [8] baselines. Our main contributions are summarized as follows:

- We propose *EpiDistill*, a novel framework that transfers scale-robust multi-view geometric priors to single-view prediction models.
- We design a *Depth-Guided Epipolar Attention* that utilizes ground-truth spatial bias along the epipolar line, providing explicit geometric supervision during the multi-view training phase.
- We introduce *Rectified Stereo Tokens*, which prevent the structural collapse of cross-view attention during single-view inference by serving as a geometric anchor in a rectified stereo setup.
- Extensive experiments demonstrate that our approach is model-agnostic and significantly improves the zero-shot metric depth performance of SoTA ViT-based models (*e.g.*, UniDepthV2, DepthPro), particularly on challenging datasets demanding strict scale alignment.

2 Related Works

2.1 Monocular Depth Estimation

Monocular depth estimation [9, 18, 42] aims to predict the depth of a scene from a single input image. Early deep learning approaches incorporated depth priors to improve performance, utilizing techniques such as ordinal regression [10] and adaptive binning to partition the depth range [1]. However, these models heavily relied on direct ground-truth supervision, which fundamentally limited their generalization capabilities across diverse datasets. To mitigate the reliance on annotated data, self-supervised methods [5, 13, 28, 40, 41, 47] were proposed. These approaches leverage novel view synthesis, utilizing nearby video frames and photometric consistency losses to jointly estimate depth and camera pose without explicit ground-truth depth.

Recently, depth foundation models trained on massive, diverse datasets have emerged. By adopting robust depth representations and range normalization techniques [2, 30, 43, 44], these models successfully aggregate training data from various domains to achieve remarkable zero-shot relative depth estimation. Building upon this success, several recent works have focused on metric depth estimation [3, 11, 25–27] by employing additional modules to predict absolute scale.

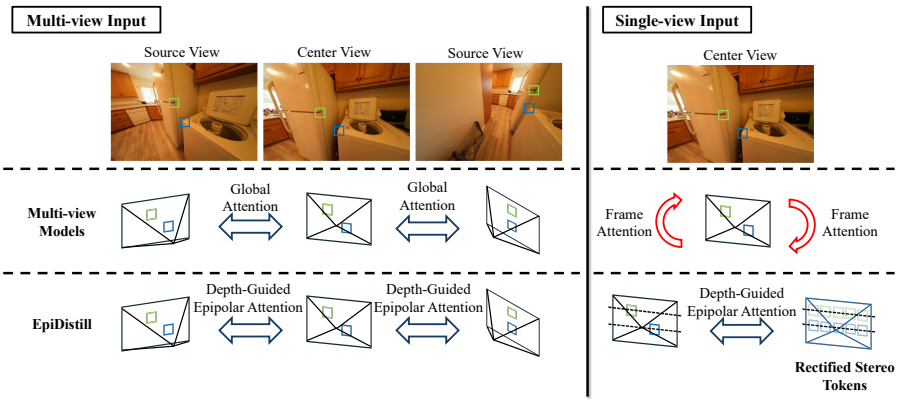

Fig. 2: Geometric ambiguity in single-view inference of multi-view models. In standard architectures, the global attention layer collapses into standard intra-frame self-attention when deprived of auxiliary views. In contrast, our proposed *EpiDistill* framework preserves the structural integrity of the cross-view attention pathways by utilizing depth-guided epipolar attention coupled with rectified stereo tokens.

Nevertheless, despite leveraging large-scale datasets, monocular metric depth estimation models inherently suffer from scale ambiguity due to their training relying on limited 2D visual cues. In contrast, our approach resolves this ambiguity by supervising the model with explicit multi-view geometric constraints during training. By distilling the capacity to reason over epipolar geometry into a single-view model via rectified stereo tokens, our framework accurately predicts metric scale during single-view inference.

2.2 Multi-view Foundation Models

Recent advances in 3D vision have introduced multi-view foundation models such as DUSt3R [39] and MASt3R [20], which utilize multiple viewpoints to jointly predict dense geometry, including point clouds and camera information. However, these methods strictly require two views as input and rely on computationally expensive iterative processes for global scene reconstruction. To overcome these computational bottlenecks, models such as VGGT [36], DepthAnything3 [21], and MapAnything [16] are proposed. They achieve globally aligned scene reconstruction and robust scale estimation through a single feed-forward pass using multi-view inputs. These architectures utilize alternating token mixing strategies, specifically interleaving intra-frame (frame) and cross-frame (global) attention layers, to establish multi-view consistency and infer scene-level scale.

Consequently, a limitation arises during single-view inference: the global attention naturally degenerates into frame attention, as shown in Fig. 2. This structural collapse discards the rich, scale-aware geometric priors learned during multi-view training, resulting in scale ambiguity similar to traditional monocular models, as shown in Fig. 1. Unlike these approaches, our proposed *EpiDis-*

till framework explicitly bridges this gap. By guiding the monocular prediction model to retain the epipolar constraints learned from multi-view training, we successfully preserve robust scale estimation capabilities without requiring additional views at inference time.

3 Preliminaries

In this section, we revisit the architecture of multi-view foundation models [16, 21, 36], which serves as the structural basis for our approach. Specifically, we formalize the tokenization and attention mechanisms responsible for learning geometric scale priors.

Multi-View Tokenization. Given a set of N images $\{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_N\}$, a Vision Transformer (ViT) [8] encoder processes each image into a sequence of spatial tokens. Let $\mathbf{F}_i \in \mathbb{R}^{L \times d}$ denote the spatial tokens for image $\mathbf{F}_i$, where L is the sequence length and d is the feature dimension. To explicitly reason about camera geometry and global scale, learnable task-specific tokens (*e.g.*, camera or scale) are appended to the spatial token sequence of each view.

Frame and Global Attention. The representational power of recent multi-view foundation models [16, 21, 36] stems from alternating attention blocks. The frame attention operates strictly within individual views, applying standard self-attention to $\mathbf{F}_i$. Conversely, the global (cross-view) attention models broad spatial relationships and structural priors across different viewpoints. Let $\mathbf{F}_{\text{global}} = [\mathbf{F}_1; \mathbf{F}_2; \ldots; \mathbf{F}_N] \in \mathbb{R}^{(N \cdot L) \times d}$ be the concatenated token sequence from all views. The global attention computes the output features by allowing tokens to attend to tokens derived from all viewpoints. With this cross-view interaction, the network inherently learns to resolve scale ambiguity by extracting multi-view geometric constraints. The final decoded outputs yield geometry predictions including depth maps and camera intrinsics.

4 EpiDistill

The overall pipeline of our proposed framework is illustrated in Fig. 3. Our approach leverages multi-view image sets (*e.g.*, $\mathbf{I}_{\text{ref}}, \mathbf{I}_{\text{src}}, \ldots$) during training to extract geometrically consistent scale priors. Specifically, we build upon SoTA ViT-based [8] depth foundation models, such as UniDepthV2 [26] and DepthPro [3]. Given overlapping multi-view images, a frozen model encoder extracts dense spatial tokens ($\mathbf{F}_{\text{ref}}$, $\mathbf{F}_{\text{src}}, \ldots$). We append a learnable scale token ($\mathbf{F}_{\text{scale}}$) to the spatial tokens of each individual view and process them through shared frame attention layers. Subsequently, the epipolar attention is applied across views to explicitly learn multi-view geometric correspondences and resolve scale ambiguity.

To enable highly accurate single-view prediction, we distill this consistent geometric information into a single-view model at the token level. To preserve the learned epipolar attention without requiring actual multi-view inputs, the single-view model substitutes the missing source views with a set of learnable

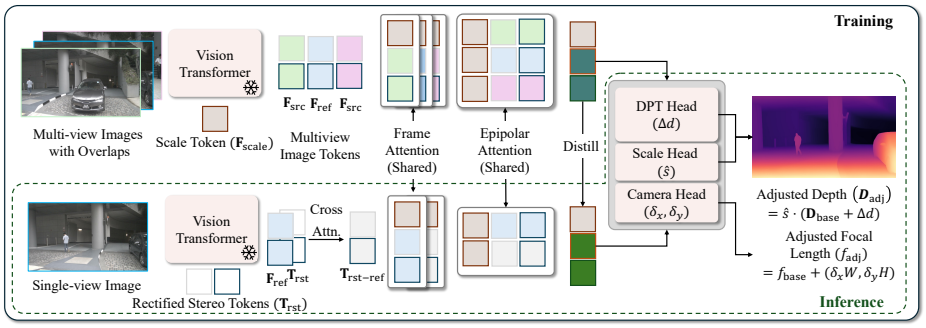

Fig. 3: Overview of the EpiDistill architecture. (Top) During training, a multi-view model extracts geometric priors via frame attention and depth-guided epipolar attention. **(Bottom)** A single-view model employs the shared layers but leverages *Rectified Stereo Tokens* as geometric guidance. Knowledge is distilled from multi-view to single-view tokens. Finally, lightweight heads predict structural shift (Δd), global scale ($\hat{s}$), and focal length (δ_x, δ_y) adjustments to refine the baseline prediction. At inference, only the single-view pipeline is utilized.

Rectified Stereo Tokens ($\mathbf{T}_{\text{rst}}$). These tokens interact with the single-view image features via a cross-attention layer to form reference-guided tokens ($\mathbf{T}_{\text{rst-ref}}$). Reference-guided tokens are then concatenated with the primary spatial tokens and processed through the shared epipolar attention layers. Finally, the geometric knowledge from the multi-view attended tokens is explicitly distilled into these single-view embeddings, yielding attention-refined spatial tokens.

Because the underlying baseline models already excel at relative depth estimation, we explicitly decouple the final depth prediction into structural and scale components. Given the original depth prediction from the foundation model as $\mathbf{D}_{\text{orig}}$, we retain this dense prediction and normalize it to obtain a scale-invariant structural depth map, $\mathbf{D}_{\text{base}} = \mathbf{D}_{\text{orig}} / (\frac{1}{N} \sum_{i=1}^{N} \mathbf{D}_{\text{orig},i})$, where N is the total number of pixels. The attention-refined spatial tokens are then fed into a lightweight Dense Prediction Transformer (DPT) [30] head to predict a dense structural shift offset, Δd. Concurrently, the scale token, enriched by both frame and epipolar attention, is passed through an MLP-based scale head to predict a metric scale factor, $\hat{s}$. The final adjusted metric depth, $\mathbf{D}_{\text{adj}}$, explicitly fuses these decoupled outputs:

$$\mathbf{D}_{\text{adj}} = \hat{s} \cdot (\mathbf{D}_{\text{base}} + \Delta d). \tag{1}$$

An auxiliary camera head resolves camera and scale ambiguities by predicting normalized focal length residuals (δ_x, δ_y) from scale tokens. These are rescaled by (W, H) to compute the adjusted focal length: $f_{\text{adj}} = f_{\text{base}} + (\delta_x W, \delta_y H)$.

4.1 Depth-Guided Epipolar Attention

We propose a novel epipolar attention network designed to effectively fuse multi-view observations and extract robust geometric priors. For a given pair of images,

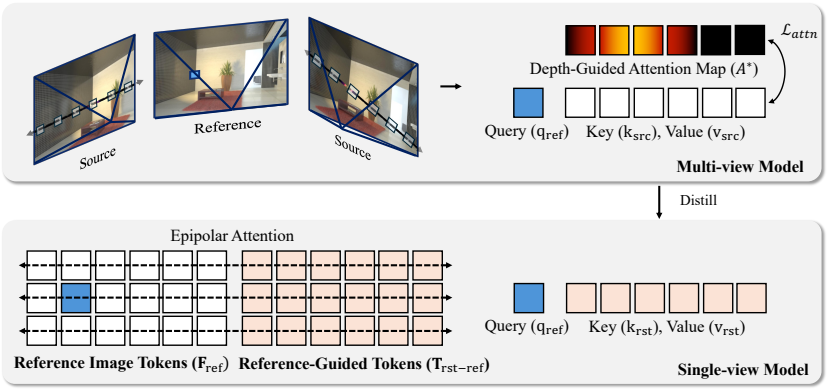

Fig. 4: Depth-Guided Epipolar Attention and Rectified Stereo Tokens. (Top) In the multi-view model, a reference query (blue) attends to points along the 2D epipolar line, supervised by $\mathcal{L}_{\text{attn}}$ to localize 3D correspondences. **(Bottom)** In single-view model, reference-guided tokens ($\text{T}_{\text{rst-ref}}$) substitute source views, simplifying the geometry into a 1D horizontal search space (dashed arrows).

depth-guided epipolar attention computes cross-attention between the tokens of the reference view and those of the source view. To enforce geometric consistency, the attention for a query token from the reference image is constrained to tokens sampled along its corresponding epipolar line in the source image.

Let $\mathbf{K}_{\text{ref}}$ and $\mathbf{K}_{\text{src}}$ be the intrinsic matrices of the reference and source view, respectively, and $[\mathbf{R}|\mathbf{t}]$ be the relative rigid transformation between the reference view and the source view. For a query pixel $\mathbf{p}_{\text{ref},i}$ in the reference view (where $i \in \{1, \ldots, N\}$ indexes the total N reference tokens) with homogeneous coordinates $\tilde{\mathbf{p}}_{\text{ref},i}$, the corresponding epipolar line $\mathbf{l}_{\text{src},i}$ in the source view is defined by the fundamental matrix $\mathbf{F} \in \mathbb{R}^{3\times3}$:

$$\mathbf{l}_{\text{src},i} = \mathbf{F}\tilde{\mathbf{p}}_{\text{ref},i} = \mathbf{K}_{\text{src}}^{-\top}[\mathbf{t}]_{\times}\mathbf{R}\mathbf{K}_{\text{ref}}^{-1}\tilde{\mathbf{p}}_{\text{ref},i}, \tag{2}$$

where $[\mathbf{t}]_{\times}$ is the skew-symmetric matrix of the translation vector $\mathbf{t}$. While standard cross-attention is applied over a set of sampled points along this line, we introduce an explicit geometric guidance mechanism using ground-truth depth during training. As illustrated in Fig. 4, by leveraging the ground-truth depth $\mathbf{D}_{\text{gt}}(\mathbf{p}_{\text{ref},i})$, the exact corresponding target point $\mathbf{p}^*_{\text{src},i}$ in the source view can be determined by back-projecting the reference pixel to 3D and projecting it onto the source image plane:

$$\mathbf{p}^*_{\text{src},i} = \pi\left(\mathbf{K}_{\text{src}}\left(\mathbf{R}\left(\mathbf{D}_{\text{gt}}(\mathbf{p}_{\text{ref},i})\mathbf{K}_{\text{ref}}^{-1}\tilde{\mathbf{p}}_{\text{ref},i}\right) + \mathbf{t}\right)\right), \tag{3}$$

where $\pi([x, y, z]^{\top}) = [x/z, y/z]^{\top}$ denotes the perspective projection operation.

To concentrate attention near the exact correspondence, we introduce a spatial Gaussian bias B_{ij} for each sampled coordinate $\mathbf{p}_{\text{src},ij}$, where $j \in \{1, \ldots, M\}$ indexes the M discrete points sampled along the epipolar line $\mathbf{l}_{\text{src},i}$:

$$B_{ij} = -\gamma\|\mathbf{p}_{\text{src},ij} - \mathbf{p}^*_{\text{src},i}\|_2^2, \tag{4}$$

where γ controls the distribution sharpness (*e.g.*, $\gamma = 50$). This bias is added directly to the raw cross-attention logits $E_{ij} = \mathbf{q}_{\text{ref},i}^{\top}\mathbf{k}_{\text{src},j}/\sqrt{d}$, yielding the depth-guided target attention map $A_{ij}^{*} = \text{Softmax}_j(E_{ij} + B_{ij})$. Here, $\mathbf{q}_{\text{ref},i}$ and $\mathbf{k}_{\text{src},j}$ denote the query token from the reference view and the key token from the source view, respectively. To distill this capability, we apply a cross-entropy loss over the predicted attention weights:

$$\mathcal{L}_{\text{attn}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} A_{ij}^{*}\log(\hat{A}_{ij}), \tag{5}$$

where $\hat{A}_{ij} = \text{Softmax}_j(E_{ij})$ represents the unguided attention predicted by the model. This aligns the attention distributions within the epipolar sampling space. Note that $\mathbf{A}^{*}$ is detached from the gradient flow to serve solely as a supervision signal. Additionally, $\mathcal{L}_{\text{attn}}$ is masked out for invalid regions due to the sparsity of the ground-truth depth, while the epipolar attention remains fully active.

4.2 Rectified Stereo Tokens

In multi-view geometry, a rectified stereo setup provides an ideal constrained environment for depth perception. By aligning the image planes, the relative camera pose simplifies to an identity rotation ($\mathbf{R} = \mathbf{I}$) and a purely horizontal translation $\mathbf{t} = [b, 0, 0]^{\top}$. Assuming identical camera intrinsics with focal length f, the fundamental matrix $\mathbf{F}$ becomes:

$$\mathbf{F} = \mathbf{K}_{\text{src}}^{-\top}[\mathbf{t}]_{\times}\mathbf{R}\mathbf{K}_{\text{ref}}^{-1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & -c \\ 0 & c & 0 \end{bmatrix}, \tag{6}$$

where $c = b/f$. For a given reference pixel $\tilde{\mathbf{p}}_{\text{ref}} = [x_0, y_0, 1]^{\top}$, the corresponding epipolar line in the source view is calculated as $\mathbf{l}_{\text{src}} = \mathbf{F}\tilde{\mathbf{p}}_{\text{ref}} = [0, -c, cy_0]^{\top}$. Since the baseline-dependent constant c cancels out in the resulting 2D line equation ($-cy + cy_0 = 0$), the geometric search space is strictly confined to $y = y_0$.

This baseline-agnostic property simplifies the epipolar attention, requiring the model to attend only to tokens along the same horizontal scanline. To preserve this structure without actual multi-view inputs, we introduce learnable *Rectified Stereo Tokens* ($\mathbf{T}_{\text{rst}}$) to serve as a constant geometric anchor. We initialize a spatial token grid $\mathbf{T}_{\text{rst}} \in \mathbb{R}^{HW\times d}$ and dynamically interpolate it to match the spatial resolution of the encoded reference features $\mathbf{F}_{\text{ref}} \in \mathbb{R}^{H'W'\times d}$. Through cross-attention, $\mathbf{T}_{\text{rst}}$ queries $\mathbf{F}_{\text{ref}}$ to embed spatial context, yielding reference-guided tokens $\mathbf{T}_{\text{rst-ref}}$. Subsequently, the model applies the rectified epipolar attention between $\mathbf{F}_{\text{ref}}$ and $\mathbf{T}_{rst-ref}$, constraining the cross-attention interaction strictly along the horizontal scanlines to explicitly simulate the rectified stereo setup.

To interpret this transformation, we visualize the high-dimensional internal representations—specifically the tokens ($\mathbf{T}_{\text{rst}}$, $\mathbf{F}_{\text{ref}}$, $\mathbf{T}_{\text{rst-ref}}$) and the epipolar attention residual map—by reducing their d-dimensional features to three principal

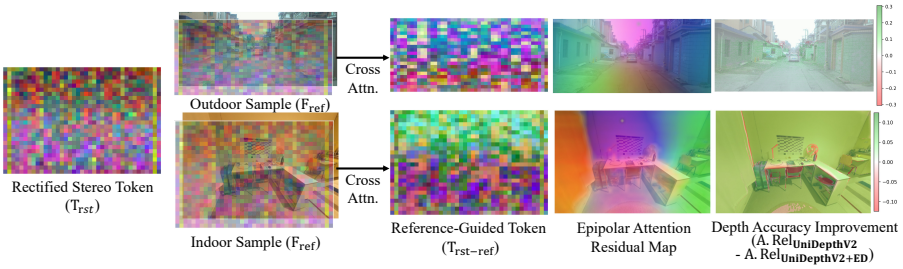

Fig. 5: Visualization of Rectified Stereo Tokens via Principal Component Analysis (PCA). The visual progression shows the PCA-projected RGB mappings of the learnable spatial tokens ($\mathbf{T}_{\text{rst}}$) undergoing cross-attention with the encoded reference features ($\mathbf{F}_{\text{ref}}$) to generate reference-guided tokens ($\mathbf{T}_{\text{rst-ref}}$). Additionally, we visualize the PCA-reduced epipolar attention residual map, alongside the resulting depth accuracy improvement map, where green regions indicate reduced A.Rel error and red indicates degradation.

components via PCA and mapping them to RGB channels. As shown in Fig. 5, the cross-attention injects scene-specific structural awareness into $\mathbf{T}_{\text{rst-ref}}$. Furthermore, we analyze the PCA-mapped epipolar attention residual, defined as the feature difference before and after applying the epipolar attention layer. This residual demonstrates that the network successfully captures semantic information, such as scene layouts. By leveraging this semantically aware geometric anchor, *EpiDistill* (ED) effectively guides the depth refinement process. This results in a significant reduction in the absolute relative error (A.Rel), as evidenced by the separate depth improvement map ($\text{A.Rel}_{\text{UniDepthV2}} - \text{A.Rel}_{\text{UniDepthV2+ED}}$), achieving accurate metric results.

4.3 Loss Functions

Our model is trained using a composite loss function that explicitly decouples structure and scale while enforcing multi-view consistency and stabilizing the distillation process. The total loss $\mathcal{L}_{\text{total}}$ is formulated as a weighted sum of its individual components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rel}} + \mathcal{L}_{\text{scale}} + \lambda_{\text{ray}}\mathcal{L}_{\text{ray}} + \lambda_{\text{distill}}\mathcal{L}_{\text{distill}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}}, \tag{7}$$

where the loss weights are empirically set to $\lambda_{\text{ray}} = 0.3$, $\lambda_{\text{distill}} = 1.0$, and $\lambda_{\text{attn}} = 0.1$.

Relative Depth Loss ($\mathcal{L}_{\text{rel}}$). To stably learn metric scale in a multi-view setting, we explicitly decouple the depth estimation into a normalized structural component and a global scale component. For the structural target, the ground-truth depth $\mathbf{D}_{\text{GT}}$ is normalized by its arithmetic mean within the valid mask region $\mathcal{M}$. The predicted relative depth $\tilde{\mathbf{D}}_{\text{rel}}$ is formed by adding a structural offset Δd to the base depth $\mathbf{D}_{\text{base}}$:

$$\tilde{\mathbf{D}}_{\text{rel}} = \mathbf{D}_{\text{base}} + \Delta d, \quad \bar{\mathbf{D}}_{\text{GT}} = \frac{\mathbf{D}_{\text{GT}}}{\frac{1}{|\mathcal{M}|}\sum_{p\in\mathcal{M}} \mathbf{D}_{\text{GT},p}}. \tag{8}$$

The structural depth loss $\mathcal{L}_{\text{depth}}$ operates in log-space using an L_1 loss:

$$\mathcal{L}_{\text{depth}} = \sqrt{\frac{1}{|\mathcal{M}|} \sum_{p \in \mathcal{M}} \left| \log \tilde{\mathbf{D}}_{\text{rel},p} - \log \bar{\mathbf{D}}_{\text{GT},p} \right|}. \tag{9}$$

Following [26], we incorporate an edge-aware loss $\mathcal{L}_{\text{edge}}$ using Sobel-filtered image gradients to preserve sharp depth discontinuities around strong object boundaries. The final relative depth loss is formulated as $\mathcal{L}_{\text{rel}} = \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{edge}}$.

Scale Loss ($\mathcal{L}_{\text{scale}}$). The metric scale $\hat{s}$ predicted by the scale head is independently supervised by the arithmetic mean of the ground-truth depth. To ensure gradient flows to the scale head, we apply an L_1 penalty in log space:

$$\mathcal{L}_{\text{scale}} = \sqrt{\left| \log \hat{s} - \log\Big(\frac{1}{|\mathcal{M}|} \sum_{p \in \mathcal{M}} \mathbf{D}_{\text{GT},p} \Big) \right|}. \tag{10}$$

Ray/Intrinsic Loss ($\mathcal{L}_{\text{ray}}$). To jointly optimize camera intrinsics, we enforce an L_1 penalty on the predicted focal length residuals (δ_x, δ_y) relative to the ground-truth focal lengths normalized by image width W and height H:

$$\mathcal{L}_{\text{ray}} = \left| \frac{f_{x,\text{base}}}{W} + \delta_x - \frac{f_{x,\text{GT}}}{W} \right| + \left| \frac{f_{y,\text{base}}}{H} + \delta_y - \frac{f_{y,\text{GT}}}{H} \right|. \tag{11}$$

Distillation Loss ($\mathcal{L}_{\text{distill}}$). To ensure stable knowledge transfer, this mechanism aligns single-view tokens $\mathbf{T}_s$ with multi-view epipolar-aggregated features $\mathbf{T}_m$ guided by the rectified stereo tokens. The loss combines a cosine similarity direction objective with a magnitude penalty over the valid token mask $\mathcal{N}$:

$$\mathcal{L}_{\text{distill}} = \frac{1}{|\mathcal{N}|} \sum_{k \in \mathcal{N}} \Big(1 - \cos\big(\mathbf{T}_{m,k}, \mathbf{T}_{s,k}\big) + 0.1 \big| \|\mathbf{T}_{m,k}\|_2 - \|\mathbf{T}_{s,k}\|_2 \big| \Big), \tag{12}$$

where $\mathcal{N}$ represents the valid token mask established by the epipolar attention, ensuring that feature alignment is focused on valid cross-view correspondences.

Attention Loss ($\mathcal{L}_{\text{attn}}$). Lastly, the attention loss $\mathcal{L}_{\text{attn}}$ explicitly guides the epipolar correspondence via cross-entropy optimization over the predicted attention weights, as previously formulated in Eq. (5).

5 Experiments

5.1 Experimental Setup

Datasets. We train our model on a multi-domain mixture of seven datasets covering indoor, outdoor, and synthetic environments: Hypersim [31], Cityscapes [6], Eden [19], ScanNet [7], ScanNet++ [45], Waymo [34], and nuScenes [4]. All selected datasets provide ground-truth depth, camera intrinsics, and 6-DoF camera poses. Using these annotations, we compute the visual overlap between frame pairs to construct training tuples consisting of a reference view and up to two source views, with triplets sampled within a ± 20 frame window.

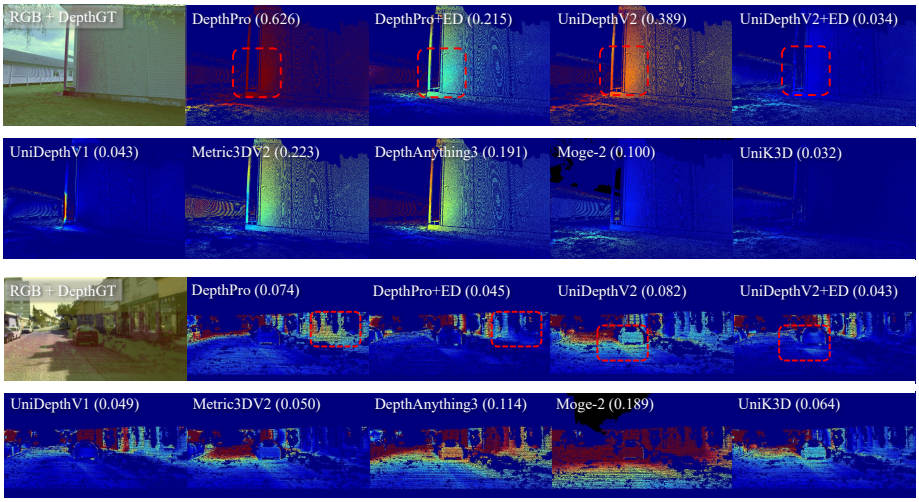

Fig. 6: Qualitative comparison on the ETH3D [32] (Top) and KITTI [12] (Bottom) datasets. The heatmaps visualize the absolute relative error (A.Rel), with the corresponding A.Rel value for each prediction reported in parentheses. ED indicates the proposed *EpiDistill* method. Note that the KITTI visualizations have been cropped by 30% on the left to improve visibility.

To ensure that the epipolar attention can reliably establish meaningful geometric correspondences, we strictly enforce a minimum visual overlap of 20% between the paired views. For the Hypersim dataset, identical sampling is applied using the exact ground-truth poses despite its discrete renderings. During training, input images are resized up to 576K pixels while preserving their aspect ratios, and the total number of training samples is capped at 50K per dataset to balance the training distribution.

Evaluation Settings. We evaluate depth prediction accuracy using standard metrics for monocular depth estimation: absolute relative error (Abs Rel), root mean squared error (RMSE), and $\delta_1 < 1.25$. Further details on these metrics can be found in the supplementary material. Crucially, our evaluation settings (*e.g.*, depth clipping caps and valid region masking) strictly follow the evaluation protocol established by UniDepthV2 [26].

Training Details. Our framework is built upon two representative Vision Transformer (ViT)-based monocular depth models: UniDepthV2 (ViT-L) [26] and DepthPro [3], utilizing their pre-trained weights. We intentionally select these two models to demonstrate the broad generalizability of *EpiDistill* across different generations and performance tiers. Specifically, UniDepthV2 represents a cutting-edge SoTA approach, while DepthPro serves as a highly influential, established baseline. By consistently improving the performance of both models, we validate that our distillation method is not overfitted to a single architecture and can be seamlessly adapted to various ViT-based frameworks.

Table 1: Zero-shot outdoor and mixed dataset evaluation. † indicates models using ground truth intrinsics during inference. Best results among models with only images as input are highlighted in **bold**, and second best are underlined. The **Rank** is calculated across all 9 models.

Method/Metrics	**KITTI** [12]		**DDAD** [13]		**DIODE** [35]		**ETH3D** [32]		**Rank** ↓
	A.Rel↓	δ_1↑	A.Rel↓	δ_1↑	A.Rel↓	δ_1↑	A.Rel↓	δ_1↑	
Metric3DV2† [15]	0.054	0.975	0.122	0.858	0.536	0.077	0.175	0.724	4.5
DepthAnything3† [21]	0.097	0.910	0.137	0.828	0.489	0.110	0.122	0.861	4.8
UniDepthV1 [27]	**0.049**	**0.980**	0.217	0.708	**0.263**	0.624	0.530	0.224	4.8
UniK3D [25]	0.123	0.940	0.156	0.859	0.457	**0.686**	0.139	0.852	4.0
MoGe-2 [38]	0.132	0.867	0.168	0.776	0.407	0.368	**0.109**	**0.911**	5.3
DepthPro [3]	0.141	0.838	0.391	0.233	0.386	0.419	0.355	0.433	8.8
+*EpiDistill*	0.123	0.875	0.348	0.353	0.357	0.462	0.354	0.393	6.4
UniDepthV2 [26]	0.080	0.945	0.144	**0.859**	0.716	0.549	0.176	0.752	4.7
+*EpiDistill*	0.074	0.951	**0.132**	0.854	0.396	0.641	0.137	0.779	**3.1**

Table 2: Zero-shot indoor dataset evaluation. † indicates models using ground truth intrinsics during inference. Best results among models with only images as input are highlighted in **bold**, and second best are underlined. The **Rank** is calculated across all 9 models.

Method/Metrics	**NYU** [33]		**Bonn** [24]		**Booster** [29]		**IBims-1** [17]		**Rank** ↓
	A.Rel↓	δ_1↑	A.Rel↓	δ_1↑	A.Rel↓	δ_1↑	A.Rel↓	δ_1↑	
Metric3DV2† [15]	0.071	0.963	0.055	0.991	0.444	0.373	0.131	0.848	4.6
DepthAnything3† [21]	0.078	0.957	0.048	0.991	0.178	0.771	0.079	0.961	**2.4**
UniDepthV1 [27]	**0.058**	**0.985**	0.062	0.986	0.468	0.324	0.446	0.133	6.0
UniK3D [25]	0.090	0.948	0.088	0.984	0.190	0.712	0.097	0.900	5.3
MoGe-2 [38]	0.077	0.945	0.211	0.540	0.262	0.605	0.114	0.862	6.1
DepthPro [3]	0.098	0.918	0.091	0.912	0.332	0.535	0.156	0.837	7.8
+*EpiDistill*	0.110	0.884	0.076	0.971	0.314	0.544	0.151	0.841	7.1
UniDepthV2 [26]	0.077	0.958	0.057	0.991	0.179	0.720	0.093	0.930	3.3
+*EpiDistill*	0.077	0.958	**0.055**	**0.991**	**0.171**	**0.727**	**0.092**	**0.935**	**2.4**

For the architecture, we freeze the backbone of the base models to extract features, and introduce a lightweight DPT head [30] to predict dense shift offsets, alongside MLP-based camera and scale heads to regress the intrinsic residual and metric scale factor, respectively. Our epipolar distillation module configures 4 frame/epipolar layers with a 37×37 rectified stereo token grid per layer.

During training, we apply random resize and center crop augmentations to simulate a diverse range of camera intrinsics. All models are trained for 100K iterations with a total batch size of 4 triplets (12 images), distributed across 8 NVIDIA H100 GPUs. The network is optimized using the AdamW [22] optimizer with a learning rate of 1×10^{-4}, momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and a weight decay of 0.01.

5.2 Zero-Shot Depth Estimation

We evaluate the zero-shot generalization of our metric depth estimation approach across diverse datasets. We compare our model against recent SoTA baselines, including Metric3Dv2 [15], UniDepthV1 [27], UniK3D [25], MoGe-2 [38], and DepthAnything3 (ViT-L, Metric) [21]. Notably, Metric3DV2 and DepthAnything3 inherently rely on ground-truth (GT) camera intrinsics during inference, whereas other methods operate solely on images.

Outdoor & Mixed-Set Evaluation. Tab. 1 evaluates geometric robustness across highly variable environments encompassing both indoor and outdoor scenes [12, 13, 32, 35]. Integrating *EpiDistill* consistently elevates baseline performance across all domains. Gains are particularly prominent on the DIODE dataset, resulting in a 33.4% and 44.7% gain in A.Rel metrics compared to DepthPro and UniDepthV2, respectively. Remarkably, UniDepthV2 + *EpiDistill* achieves the best overall rank (3.1), outperforming even SoTA models reliant on GT intrinsics. As visually corroborated in highlighted regions of Fig. 6, *EpiDistill* uniquely empowers both baselines to comprehensively resolve severe scale ambiguities and recover sharp structural details, yielding robust and globally consistent metric depth.

Indoor Evaluation. Tab. 2 reports quantitative results across four constrained indoor domains [17, 24, 29, 33]. *EpiDistill* brings stable improvements to both DepthPro [3] and UniDepthV2 [26]. On the Booster dataset, our framework reduces the A.Rel error by 5.4% and 4.5% for DepthPro and UniDepthV2, respectively. Ultimately, UniDepthV2 + *EpiDistill* achieves an average rank (2.4) matching the GT intrinsic-dependent DepthAnything3. While *EpiDistill* primarily targets the severe scale ambiguities of unconstrained scenes, these results confirm that it simultaneously preserves and enhances accuracy in indoor environments where scale variations are less dominant.

5.3 Generalization Across Diverse Scales

To investigate our framework's robustness against extreme scale ambiguity, we evaluate scale generalization on the DepthPerturb dataset [23], utilizing Dolly Zoom sequences. A Dolly Zoom effect occurs when the camera translates along the optical axis while simultaneously adjusting the focal length to keep the main subject's 2D size constant on the image plane. This configuration uniquely challenges a model's ability to disentangle metric depth from camera intrinsics, revealing whether it genuinely understands 3D geometry.

Since the dataset lacks ground truth camera information, we assess this disentanglement by reporting standard depth metrics alongside the Pearson Correlation (PCorr) for intrinsic evaluation. In a Dolly Zoom setup, the focal length is deliberately adjusted linearly to counteract the camera translation. Therefore, PCorr serves as a reliable indicator of the model's ability to capture this linear geometric trend. Formally, for a video of T frames, let t denote the frame index and $\hat{f}_t$ the predicted focal length. To measure temporal linearity, PCorr is com-

puted as $|\text{PCorr}| = |\sum(t-\bar{t})(\hat{f}_t-\bar{\hat{f}})|/\sqrt{\sum(t-\bar{t})^2\sum(\hat{f}_t-\bar{\hat{f}})^2}$, where $\bar{t}$ and $\bar{\hat{f}}$ are their respective means over the sequence.

As reported in Tab. 3, we evaluate three intrinsic-predicting baselines [3,26,27]. Baseline models fail to perceive dynamic depth changes and struggle to predict the linearly varying intrinsics. In contrast, our *EpiDistill* framework demonstrates exceptional generalization, significantly reducing depth errors (e.g., A.Rel drops from 0.241 to 0.167 for UniDepthV2, and 0.248 to 0.178 for DepthPro). Furthermore, it yields an improved focal length prediction trajectory, characterized by maximized PCorr scores. This confirms that transferring multi-view epipolar geometry equips the single-view model with the robust, intrinsic-aware scale priors necessary to resolve depth-intrinsic ambiguity.

Table 3: Scale generalization evaluation on the DepthPerturb dataset [23]. *EpiDistill* demonstrates superior disentanglement of metric depth and focal length.

Method	Depth			Intrinsic
	A.Rel ↓	RMS ↓	δ_1 ↑	|PCorr| ↑
UniDepthV1 [27]	0.432	0.957	0.350	0.704
DepthPro [3]	0.248	0.596	0.598	0.906
+ *EpiDistill*	0.178	**0.498**	0.720	0.911
UniDepthV2 [26]	0.241	0.853	0.689	0.955
+ *EpiDistill*	**0.167**	0.694	**0.843**	**0.965**

5.4 Ablation Study

We conduct an ablation study to validate the individual and synergistic contributions of our proposed modules, namely Depth-Guided Epipolar Attention (DGEA) and Rectified Stereo Tokens (RST), using the UniDepthV2 model. To ensure a comprehensive evaluation, we report the averaged performance metrics across the KITTI [12] and ETH3D [32] datasets. The fine-tuned baseline refers to directly fine-tuning the decoder with a learning rate of 1×10^{-6}.

We first report the performance of the fine-tuned UniDepthV2 trained on our dataset. As an additional baseline, we employ the standard global cross-attention mechanism used in recent multi-view foundation models [21, 36], which degenerates into intra-frame self-attention during single-view inference. Note that we omit an RST-only ablation variant, as the regularized sparse tokens inherently rely on the epipolar attention layer to establish and process geometric correspondences. Computational costs (FLOPs and latency) are measured on a single NVIDIA RTX A6000 GPU.

Depth-Guided Epipolar Attention (DGEA). The model is trained with the DGEA layer but inferred without RST. Consequently, during single-view inference, the reference image tokens serve simultaneously as query, key, and value, functioning similarly to standard frame attention. Tab. 4 shows consistent improvements across all metrics with negligible computational overhead. This indicates that DGEA effectively enforces explicit spatial constraints during multi-view training, embedding a geometric prior that indirectly benefits single-view inference compared to global attention.

EpiDistill (Full Framework). Integrating both DGEA and RST yields the most significant performance boost. RST preserves the cross-view attention path-

Table 4: Ablation study on the effectiveness of our proposed modules. Depth-Guided Epipolar Attention, Rectified Stereo Tokens. Best results are highlighted in **bold**.

Components		Computation Cost		Metrics		
DGEA	RST	Memory	Inference time	A.Rel ↓	RMS ↓	δ_1 ↑
UniDepthV2 (Fine-tuned)		57.87M	131ms	0.124	2.721	0.840
UniDepthV2 (Global Attn.)		154.65M	180ms	0.119	2.649	0.858
✓		155.65M	190ms	0.113	2.419	0.866
✓	✓	180.98M	218ms	**0.101**	**2.206**	**0.870**
Multi-view Model		180.98M	396ms	0.100	2.102	0.880

ways optimized by DGEA during multi-view training. By maintaining this continuous flow of geometry-aware information, our complete *EpiDistill* framework explicitly mitigates scale ambiguity, achieving the highest δ_1 accuracy and the lowest errors among single-view configurations, including a 16.7% reduction in RMSE compared to the baseline.

Multi-view Upper Bound. Finally, we evaluate the multi-view model to establish a performance upper bound. As shown in Tab. 4, while the multi-view model requires identical memory, processing three input views nearly doubles the inference time to 396 ms. Notably, our single-view *EpiDistill* closely approaches this multi-view upper bound with only marginal differences in overall metrics. This firmly demonstrates that *EpiDistill* successfully distills multi-view geometric knowledge into a single-view predictor, effectively achieving multi-view-level accuracy without the computational overhead.

6 Conclusion

In this paper, we presented *EpiDistill*, a novel geometric distillation framework designed to tackle the persistent challenge of scale ambiguity in monocular metric depth estimation. While recent multi-view foundation models exhibit robust scale perception, they inherently suffer from structural attention ambiguity when restricted to single-image inference. To bridge this gap, our approach transfers the geometry-aware scale priors from a multi-view to a single-view model.

At the core of our methodology are two key innovations: the *Depth-Guided Epipolar Attention*, which enforces strict spatial constraints during multi-view training, and *Rectified Stereo Tokens*, which implicitly construct a geometric anchor to preserve cross-attention pathways during single-view inference. Extensive zero-shot evaluations demonstrate that our model-agnostic approach significantly enhances the performance of SoTA ViT-based baselines. By effectively decoupling and retaining multi-view scale knowledge, *EpiDistill* exhibits remarkable robustness across diverse outdoor and indoor environments, and proves particularly resilient against extreme depth-intrinsic entanglements, as evidenced by our scale generalization analysis.

References

1. Bhat, S.F., Alhashim, I., Wonka, P.: Adabins: Depth estimation using adaptive bins. In: CVPR (2021)
2. Birkl, R., Wofk, D., Müller, M.: Midas v3. 1–a model zoo for robust monocular relative depth estimation. arXiv preprint arXiv:2307.14460 (2023)
3. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: ICLR (2025)
4. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020)
5. Choi, H., Lee, H., Kim, S., Kim, S., Kim, S., Sohn, K., Min, D.: Adaptive confidence thresholding for monocular depth estimation. In: ICCV (2021)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
7. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR (2017)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
9. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. In: NeurIPS (2014)
10. Fu, H., Gong, M., Wang, C., Batmanghelich, K., Tao, D.: Deep ordinal regression network for monocular depth estimation. In: CVPR (2018)
11. Ganesan, G.C., Guo, Y., Ren, L., Liu., X.: Unidac: Universal metric depth estimation for any camera. In: CVPR (2026)
12. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. IJRR (2013)
13. Guizilini, V., Ambrus, R., Pillai, S., Raventos, A., Gaidon, A.: 3d packing for self-supervised monocular depth estimation. In: CVPR (2020)
14. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: ICCV (2023)
15. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. TPAMI (2024)
16. Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N., et al.: Mapanything: Universal feed-forward metric 3d reconstruction. In: 3DV (2026)
17. Koch, T., Liebel, L., Körner, M., Fraundorfer, F.: Comparison of monocular depth estimation methods using geometrically relevant metrics on the ibims-1 dataset. CVIU (2020)
18. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: 3DV (2016)
19. Le, H.A., Mensink, T., Das, P., Karaoglu, S., Gevers, T.: Eden: Multimodal synthetic dataset of enclosed garden scenes. In: WACV (2021)
20. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with mast3r. In: ECCV (2024)

21. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. In: ICLR (2026)
22. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: ICLR (2019)
23. Nugent, J., Wu, S., Ma, Z., Han, B., Parakh, M., Joshi, A., Mei, L., Raistrick, A., Li, X., Deng, J.: Evaluating robustness of monocular depth estimation with procedural scene perturbations. arXiv preprint arXiv:2507.00981 (2025)
24. Palazzolo, E., Behley, J., Lottes, P., Giguère, P., Stachniss, C.: ReFusion: 3D Reconstruction in Dynamic Environments for RGB-D Cameras Exploiting Residuals. In: IROS (2019)
25. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Van Gool, L.: Unik3d: Universal camera monocular 3d estimation. In: CVPR (2025)
26. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. TPAMI (2025)
27. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: CVPR (2024)
28. Poggi, M., Aleotti, F., Tosi, F., Mattoccia, S.: On the uncertainty of self-supervised monocular depth estimation. In: CVPR (2020)
29. Ramirez, P.Z., Tosi, F., Poggi, M., Salti, S., Mattoccia, S., Di Stefano, L.: Open challenges in deep stereo: the booster dataset. In: CVPR (2022)
30. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021)
31. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021)
32. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: CVPR (2017)
33. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
34. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020)
35. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. CoRR (2019)
36. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025)
37. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: CVPR (2025)
38. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. In: NeurIPS (2025)
39. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024)
40. Watson, J., Firman, M., Brostow, G.J., Turmukhambetov, D.: Self-supervised monocular depth hints. In: ICCV (2019)
41. Wong, A., Fei, X., Tsuei, S., Soatto, S.: Unsupervised depth completion from visual inertial odometry. RAL (2020)

42. Xu, D., Ricci, E., Ouyang, W., Wang, X., Sebe, N.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: CVPR (2017)
43. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
44. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS (2024)
45. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023)
46. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: ICCV (2023)
47. Zhu, S., Liu, X.: Revisit self-supervision with local structure-from-motion. In: ECCV (2024)