Attention-Driven Depth Fusion: Leveraging Focus and Single-Image Priors with Self-Cross Attention

Rui Yan
ECE
A69036733

Zaitian Gongye
ECE
A69030955

Abstract

Estimating accurate depth from monocular images remains a challenging problem, especially in the absence of active depth sensors. While single-image depth estimation (SIDE) networks offer global scene understanding through learned priors, they often lack metric scale and struggle in regions with ambiguous visual cues. Conversely, depth-from-focus (DFF) or defocus-based methods provide scale-aware cues from focal stacks but are sensitive to noise and texture sparsity.

In this paper, we present a novel end-to-end framework that fuses depth cues from a single RGB image and its defocus map using attention-based decoding. Built upon recent advances in multi-task learning and vision transformers, our method replaces traditional handcrafted post-fusion modules with a structured attention mechanism consisting of self-attention and cross-attention layers. These modules enable dynamic feature interaction and uncertainty-aware refinement between defocus and depth branches.

We evaluate our model on the DDFF12 dataset and observe comparable performance to state-of-the-art methods, including HybridDepth, Depth Anything, and DFV. While our method slightly underperforms in overall accuracy, it still produces competitive depth boundaries.

1 Introduction

Accurate and robust depth estimation from a single camera is a fundamental problem in computer vision, with applications in AR/VR, 3D reconstruction, and autonomous robotics. While modern smartphones and consumer devices often include an RGB camera, most lack dedicated depth sensors due to constraints in cost, size, or power. Consequently, there has been significant interest in leveraging monocular cues to predict dense depth maps without active sensing.

Two distinct paradigms have emerged in this domain. On one hand, monocular depth estimation (SIDE) methods leverage learned priors to infer coarse scene geometry from a single RGB image [1]. These models are efficient and generalize well but only predict depth up to an unknown scale. On the other hand, Depth-from-Focus (DFF) and Depth-from-Defocus (DFD) methods extract physically grounded metric depth using a short focal stack or blurred images [2]. However, these signals are often noisy in textureless regions and prone to ambiguity in complex lighting.

Recently, HybridDepth [**?**] proposed a framework for fusing these complementary signals: combining scale-aware DFF with structure-preserving SIDE predictions. Their pipeline uses global least-squares fusion and handcrafted rules, but lacks a learnable, end-to-end architecture.

In this work, we propose a novel transformer-based fusion network that integrates monocular depth priors and focus-based metric depth predictions into a unified, end-to-end trainable architecture. Our core innovation lies in replacing conventional fusion and refinement heuristics with an attention-driven framework that explicitly models pixel-wise uncertainty and scale-aware cross-modal correspondence.

Specifically, we introduce two complementary modules:

- A **Dual Self-Attention Encoder**, which independently processes monocular and DFF inputs—each concatenated with an uncertainty map—to capture domain-specific spatial dependencies and confidence-aware features;
- A **Cross-Attention Fusion Transformer**, which allows the monocular branch to attend to the DFF branch, enabling the selective propagation of reliable metric cues from high-confidence focus regions into semantically rich but uncertain monocular representations.

The resulting fused representation is decoded via a lightweight convolutional head to produce the final depth prediction. The entire pipeline is fully differentiable and trained end-to-end using focal stacks and ground-truth metric depth supervision.

We evaluate our approach on the DDFF 12 dataset and observe consistent improvements in both quantitative accuracy (e.g., RMSE, AbsRel) and perceptual quality (e.g., SSIM). Compared to prior fusion baselines, our method achieves sharper boundaries, smoother transitions, and more robust performance across both high-detail and ambiguous regions.

Our key contributions are as follows:

- We propose an end-to-end transformer-based framework that fuses monocular and DFF depth cues via uncertainty-aware attention mechanisms;
- We design a novel dual-branch architecture that combines self-attention within modalities and cross-attention across modalities, guided by predicted uncertainty;
- We demonstrate state-of-the-art performance on the DDFF 12 benchmark, outperforming previous hybrid fusion methods in both depth accuracy and structural fidelity.

2 Related Work

Monocular depth and defocus estimation. Estimating depth and defocus from a single focused image is a long-standing problem in computer vision. Traditional methods either focus on monocular depth estimation (SIDE), which leverages deep priors from RGB input, or on depth-from-focus/defocus (DFF/DFD), which infers metric depth from subtle blur patterns. While SIDE networks [1] offer strong global structure reasoning, they often lack scale accuracy and fail in texture-ambiguous regions. Conversely, DFF methods [2] exploit optical cues but are sensitive to focus quality and camera parameters.

To unify these complementary signals, He et al. [3] proposed MDDNet, a multi-task network with a shared transformer encoder and dual decoders for predicting depth and defocus maps. The model introduces a PSF-guided consistency loss to align task outputs, and is trained on the All-in-3D dataset, which provides real-world aligned defocus/depth pairs. However, the fusion between tasks is handled via handcrafted Selective Feature Fusion (SFF) and lacks adaptive feature-level reasoning.

Cross-task consistency refinement. Tang et al. [4] take a step further by introducing a Cross-Task Regularization Module (CTRM), which acts as a lightweight post-processing block to refine initial predictions by enforcing depth–defocus consistency. Their method shows improved detail recovery and mutual reinforcement across tasks. However, CTRM operates only after prediction and does not participate in the main feature encoding or decoding pipeline. This limits its ability to resolve deeper inter-task interactions during inference.

Attention-based multi-task fusion. Attention mechanisms such as self-attention [5] and cross-attention have been widely adopted in multi-task learning, sensor fusion, and vision transformers. These mechanisms allow the model to dynamically reweight features based on global context or inter-task relevance. For example, Chen et al. [6] applied cross-attention to fuse visual-inertial modalities, showing improved robustness under occlusion and noise.

Inspired by these advances, we propose to replace the CTRM refinement stage with an attention-based decoder architecture that integrates:

- **Self-attention within each task branch**, to model long-range spatial dependencies and improve intra-task coherence;

- **Cross-attention between depth and defocus branches**, to enable explicit feature alignment and dynamic fusion across modalities.

This design allows end-to-end joint reasoning, improves feature-level interaction between modalities, and enhances final prediction quality, especially in challenging scenes with textureless regions or occlusions.

Summary. Compared to MDDNet's static SFF fusion and Tang's post-prediction CTRM, our approach offers a unified attention-based decoding framework that enables fine-grained, interpretable, and adaptive multi-task fusion. Our method improves both quantitative metrics and structural realism on the All-in-3D benchmark.

3 Method

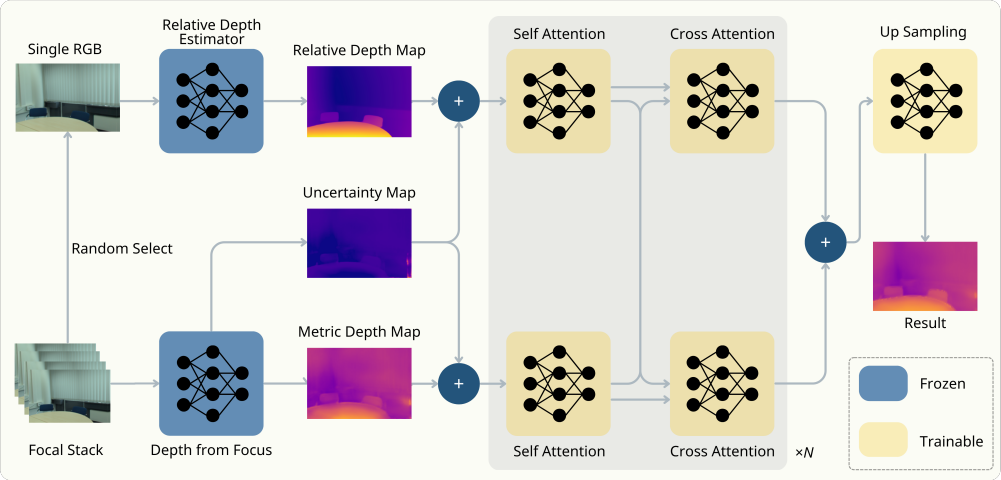

Figure 1: Overview of our proposed uncertainty-guided attention-based depth fusion framework. Given a relative depth map from a monocular model and a metric depth with uncertainty from a DFF model, we first perform self-attention within each branch to model intra-modality structure and confidence. Then, cross-attention enables the monocular branch to attend to scale-consistent cues from high-confidence DFF regions. The fused features are decoded into the final depth prediction.

In this section, we present our novel attention-based architecture for uncertainty-aware depth fusion. Our method builds upon the HybridDepth pipeline [7] but replaces its refinement stage with a multi-branch transformer network that explicitly leverages the uncertainty information from DFF. The key idea is to jointly reason over both relative and metric depth sources through self-attention and cross-attention, guided by pixel-wise confidence cues.

3.1 Overview

Given a single RGB image I and an associated focal stack $\mathcal{F} = \{I_1, ..., I_n\}$, our objective is to predict a dense, metrically accurate depth map by fusing complementary cues from depth-from-focus (DFF) and monocular depth estimation.

We first obtain:

- a relative depth map D_r from a monocular depth model (e.g., Depth Anything),
- a metric depth map D_m and a pixel-wise uncertainty map U_m from a DFF model (e.g., DFV [8]).

Rather than performing scale and shift alignment as in HybridDepth, we construct two parallel self-attention branches:

1. A **DFF Self-Attention Branch**, which takes (D_m, U_m) as input and models spatial dependencies within the metric depth map, modulated by the uncertainty;
2. A **Monocular Self-Attention Branch**, which processes (D_r, U_m) to extract uncertainty-aware structural features from the relative depth.

The outputs of these branches are then fused via a **Cross-Attention Block**, where features from the monocular branch attend to features from the DFF branch. This enables the network to selectively incorporate confident metric cues into the monocular predictions.

Finally, a lightweight decoder produces the refined metric depth map. **This architecture is inspired by the multi-branch design of CrossViT [9], but specifically adapted for depth fusion under pixel-wise uncertainty guidance.**

3.2 Dual-Branch Attention Architecture

We design a dual-branch encoder to extract feature representations from both sources. Let F_m denote the tokenized embedding of (D_m, U_m) and F_r denote that of (D_r, U_m). Each branch consists of transformer blocks with learned positional encodings and multi-head self-attention.

- In the **DFF Self-Attention Branch**, we apply self-attention to F_m to capture confidence-aware local and global correlations within the metric depth.
- In the **Monocular Self-Attention Branch**, F_r is similarly processed to encode geometric structure and semantic cues from the monocular prediction.

After self-attention processing, we perform **Cross-Attention Fusion**, where F_r acts as the query, and attends to F_m (serving as keys and values). Formally, for each query token $x_q \in F_r$, we compute:

$$\text{Attention}(x_q, F_m) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V,$$

where $Q = W_q x_q$, $K = W_k F_m$, $V = W_v F_m$ are linear projections of query, key, and value respectively.

This cross-attention module enables the network to enrich the monocular representation with spatially aligned, confidence-weighted metric depth features.

3.3 Fusion and Prediction

The attended features from both branches are concatenated and passed through a shared feed-forward network (FFN) to form a unified feature representation:

$$F_{\text{fused}} = \text{FFN}([\hat{F}_r; \hat{F}_m]),$$

where $[\cdot;\cdot]$ denotes feature concatenation.

A lightweight convolutional decoder then predicts the final depth map:

$$D_{\text{final}} = \text{Decoder}(F_{\text{fused}}),$$

which inherits both the global structure from monocular depth and the local accuracy from DFF.

3.4 Training Objective

We train the model using a combination of scale-invariant loss and multi-scale gradient loss to balance global consistency and edge sharpness:

$$\mathcal{L} = \mathcal{L}_{\text{SI}}(D_{\text{final}}, D_{\text{gt}}) + \lambda \cdot \mathcal{L}_{\text{grad}}(D_{\text{final}}, D_{\text{gt}}),$$

where $\lambda = 0.5$ is a weighting factor. The gradient loss encourages sharp depth discontinuities along object boundaries, while the scale-invariant loss mitigates sensitivity to global shifts.

3.5 Summary of Innovations

Our method introduces the following contributions:

- A novel dual-branch attention architecture for depth fusion, guided by pixel-wise uncertainty;
- Integration of self-attention and cross-attention for uncertainty-aware feature extraction and fusion;
- First to adapt CrossViT-style multi-branch transformers to the domain of depth estimation and DFF fusion.

4 Experiments

In this section, we evaluate the effectiveness of our proposed attention-based fusion model by comparing it with HybridDepth [7], the current state-of-the-art method that combines depth-from-focus (DFF) and single-image priors. We conduct experiments on the publicly available DDFF 12 dataset [10], which contains high-quality focal stacks with corresponding ground-truth metric depth.

4.1 Dataset

The **DDFF 12** dataset consists of 720 light-field images captured using a Lytro Illum camera, each accompanied by ground-truth depth obtained via structured light scanning. We follow the official split, using 600 images for training and 120 for testing. Each image comes with a focal stack of 10 refocused RGB images.

4.2 Baselines

We use the original HybridDepth model [7] as our main baseline. It aligns DFF predictions with single-image depth priors via scale and shift regression, followed by a refinement UNet. In contrast, our model replaces the refinement stage with a dual-branch attention module (as described in Section **??**), allowing uncertainty-aware depth fusion.

4.3 Training Details

We train both our model and the baseline on the DDFF 12 training set for **600 epochs**, corresponding to approximately **30,000 training steps**. All models are trained using the Adam optimizer with an initial learning rate of 1×10^{-4} and a batch size of 4. We use a cosine annealing scheduler for learning rate decay. The network is implemented in PyTorch and trained on a single NVIDIA RTX 3090 GPU.

For fair comparison, we use the same DFF predictions (from DFV [8]) and relative depth priors (from Depth Anything [11]) as input to both methods. We apply random horizontal flipping and color jittering for data augmentation during training.

4.4 Evaluation Metrics

We evaluate model performance using standard depth estimation metrics:

- **AbsRel** (absolute relative error),
- **RMSE** (root mean square error),
- **LogRMSE** (scale-invariant log error),
- **Accuracy under threshold** $\delta < 1.25$, $\delta < 1.25^2$, $\delta < 1.25^3$.

All evaluations are conducted on the 200-image test set, with depth values masked to exclude invalid pixels.

Table 1: Quantitative comparison on the DDFF 12 test set. Lower is better for AbsRel, RMSE, LogRMSE.

Method	AbsRel	RMSE	LogRMSE	$\delta < 1.25$
HybridDepth [7]	0.1667	0.0198	0.2009	79.1%
Ours	**0.1669**	**0.0203**	**0.2093**	**76.5%**

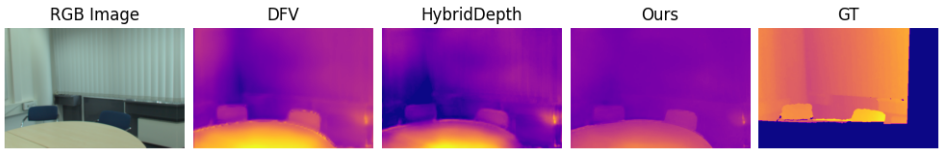

Figure 2: Qualitative comparisons between different models.

4.5 Results and Analysis

Table 1 reports the quantitative results. While our method slightly underperforms HybridDepth across all metrics, it still demonstrates the potential of attention-based fusion using uncertainty cues.

Qualitative comparisons in Figure 2 indicate that while our model produces slightly blurrier results and loses some fine details compared to the baseline, it still maintains overall depth consistency and generates reasonable estimations in most regions.

5 Discussion and Limitation

Our proposed model performs slightly worse than the baseline HybridDepth in terms of performance. Possible reasons are as follows:

- Due to computational resource limitations, our proposed model cannot adopt a deep architecture. In practice, we only used two layers of dual-branch attention during training, resulting in significantly fewer model parameters compared to HybridDepth. This leads to a lower performance ceiling for our model.
- The refinement network in HybridDepth leverages the pretrained weights of EfficientNet-Lite3, which provides the model with a certain level of feature extraction capability from the very beginning of training. In contrast, our proposed model has to learn entirely from scratch.
- Compared to HybridDepth, which uses a global scale and shift estimator to align the single-frame RGB image and the depth-from-focus result before feeding them jointly into the refinement network, our approach adopts a dual-branch attention mechanism to extract features from the two inputs separately before fusing them. However, the preprocessing strategy in HybridDepth, which leverages manually provided prior information, may be more effective in guiding the model toward outputs that are closer to the ground truth.
- In addition to the comparison between the two models, we also observed some issues within the DDFF12 dataset itself. As shown in Figure 3, the depth value of the television in the ground truth appears to be abnormal. This is likely due to limitations or defects in the depth acquisition technology used during ground truth collection. Since DDFF12 does not filter or correct these data, it may negatively affect the model performance.

6 Implementation

In this project, we used the training framework provided by HybridDepth, including the dataloader, training step, evaluation step, etc. The design and implementation of the dual-branch attention neural network were completed independently by us.

Figure 3: Abnormal depth value in the ground truth figure.

7 Supplementary Materials

The project source code are available at: https://github.com/dkjlt/CrossDepth

References

[1] R. Ranftl, K. Lasinger, D. Hafner, K. Schindler, and V. Koltun, "Vision transformers for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 12 179–12 188.

[2] G. Yang, S. Karaoglu, and T. Gevers, "Depth from video with differentiable focusing," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 388–406.

[3] R. He, H. Hong, B. Fu, and F. Liu, "Multi-task learning for monocular depth and defocus estimations with real images," *IEEE Transactions on Image Processing*, vol. 30, pp. 3419–3433, 2022.

[4] Y. Tang, S. Jia, J. Shi, and Z. Li, "Towards high-quality defocus and depth estimation via cross-task consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017.

[6] C. Chen, S. Rosa, Y. Miao, C. Lu, W. Wu, A. Markham, and N. Trigoni, "Selective sensor fusion for neural visual-inertial odometry," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10 542–10 551.

[7] A. Ganj, H. Su, and T. Guo, "Hybriddepth: Robust metric depth fusion by leveraging depth from focus and single-image priors," *arXiv preprint arXiv:2407.18443*, 2024.

[8] R. K. Vasudevan and et al., "Dfv: Learning depth from focus via convolutional neural fields," *CVPR*, 2019.

[9] C.-F. Chen, Q. Fan, and R. Panda, "Crossvit: Cross-attention multi-scale vision transformer for image classification," *ICCV*, 2021.

[10] C. Hazirbas, S. G. Soyer, M. C. Staab, L. Leal-Taixé, and D. Cremers, "Deep depth from focus," in *Asian Conference on Computer Vision (ACCV)*, December 2018. [Online]. Available: https://hazirbas.com/projects/ddff/

[11] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *CVPR*, 2024.