MS360: A Multi-Scale Feature Fusion Framework for 360 Monocular Depth Estimation

Payal Mohadikar
pvm6b8@umystem.edu
University of Missouri
Columbia, Missouri, USA

Chuanmao Fan
cf7b6@umsystem.edu
University of Missouri
Columbia, Missouri, USA

Ye Duan
duan@clemson.edu
Clemson University
Clemson, South Carolina, USA

ABSTRACT

Panorama images are popularly used for comprehensive scene understanding due to their integrated field of view. To overcome the spherical image distortions observed in commonly used Equirectangular Projection (ERP) 360-format images, the existing 360 monocular deep learning-based depth estimation networks propose using distortion-free tangent patch images projected from ERP to predict perspective depths which are merged to get the final ERP depth map. These methods show improved performance over previous methods; however, they produce depth maps that are inconsistent, and uneven, have merging artifacts, and miss fine structure details due to the missing holistic contextual information in the learned local tangent patch image features. To address this problem, we propose a novel multi-scale 360 monocular depth estimation framework, MS360, which focuses on guiding the local tangent perspective image features with coarse integrated image features. Specifically, our method first extracts coarse comprehensive features with perspective tangent patches from downsampled ERP as input to the coarse UNet structure. Secondly, we use a fine branch network to capture local geometric information using perspective tangent images from high-resolution ERP. Furthermore, we present a Multi-Scale Feature Fusion (MSFF) bottleneck module to fuse and guide the fine local features with coarse holistic features via an attention mechanism. Lastly, we predict a low-resolution depth map using coarse features and a final high-resolution depth map using coarse-guided fine image features as input to the coarse and fine decoder networks. Our method greatly reduces the discrepancies, and local patch merging artifacts in the depth maps. Performed experiments on multiple real-world depth estimation benchmark datasets show that our network outperforms the existing models both quantitatively and qualitatively while producing smooth and high-quality depth maps.

CCS CONCEPTS

• **Computing methodologies → Scene understanding**.

KEYWORDS

360, Depth, Monocular

GI '24, June 03–06, 2024, Halifax, NS, Canada

ACM ISBN 978-1-4503-XXXX-X/18/06
https://doi.org/XXXXXXX.XXXXXXX

ACM Reference Format:
Payal Mohadikar, Chuanmao Fan, and Ye Duan. 2024. MS360: A Multi-Scale Feature Fusion Framework for 360 Monocular Depth Estimation. In *Proceedings of Graphics Interface (GI '24).* ACM, New York, NY, USA, 11 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Over the years much research has been carried out on perceiving and analyzing scenes using computer vision tasks like depth estimation, semantic segmentation, object detection, etc. In the last few decades due to the boost in autonomous driving, scene reconstruction, virtual reality, and augmented reality applications identifying depth for 3D scene understanding has been the fundamental and most dominant research topic. Early on researchers used a stereo-matching-based approach [17, 22, 41] to predict depths from multiple perspective images. Eventually, more feasible single-perspective image-based depth estimation methods [3, 16, 24, 26, 40] became popular with the development of deep learning technology. However, the perspective images used in these methods provide a limited field-of-view scene content. Therefore, with the advancement of spherical cameras, 360 images started being increasingly utilized as they provide a wide field of view scene representation. Unfortunately, the widely used 360 image format Equirectangular Projection (ERP) contains distortions due to the spherical singularity at the top and bottom latitude regions. Hence, depth estimation task using a single 350-image input requires distinctive remodeling of the conventional deep learning networks.

Many deep learning models have been proposed over time to mitigate the ERP distortion problem for depth estimation tasks. For example, methods in [5, 8–11, 14, 15, 20, 32, 33, 36, 43–45] either use spherical domain filters or distortion aware convolution neural networks (CNN) to explicitly learn the distortion. While methods like BiFuse [38], UniFuse [21] and BiFuse ++ [39] propose using a combination of distortion-less cube map and ERP for better performance. HoHoNet [35] method used panorama image latent horizontal features to estimate depths. Recently, some neural networks in [7, 13, 25, 30, 31, 34] showed that using distortion-less tangent images to estimate perspective depth maps which when merged to ERP show better results. Although this approach produces better results than the previous works, they still disregard learning the ERP input integrated contextual information resulting in inconsistent, uneven, local patch merging artifacts and poor structure detailed depth maps.

This paper proposes a novel multi-scale 360 monocular depth estimation network, MS360, as shown in Figure 1 to mitigate the observed discrepancies in the SOTA method-produced depth maps. Our framework introduces a coarse and fine dual-branch network to learn integrated and local image features respectively which are

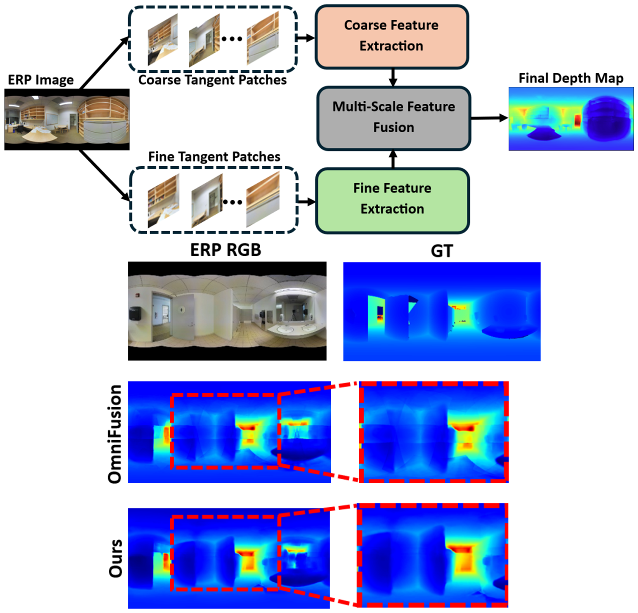

Figure 1: (Top) An overview of our model, MS360. We use a coarse-fine depth estimation approach that fuses coarse and fine perspective tangent patch (projected from ERP input) image features using a Multi-Scale Feature Fusion (MSFF) attention-based module to guide local patches with integrated coarse features for improved perspective depth map aggregation to ERP format. (Bottom) Our model produces smooth, consistent, and detailed structure depth maps compared to State-Of-The-Art (SOTA) OmniFusion [25].

then merged using a Multi-Scale Feature Fusion (MSFF) attention-based network to produce a more refined and consistent ERP depth map at the output. First, our model takes low and high-resolution ERP images and projects them to coarse and fine-level perspective tangent patches respectively using gnomonic projection (See Figure 3). These multi-scale tangent patches and their 3D geometric information are given to the coarse and fine encoders to extract comprehensive and fine-level image features. Secondly, to guide the local features with the holistic feature information we fuse the obtained bottleneck coarse and fine-level image features using the attention mechanism. Finally, these coarse and updated fine image features are given to the coarse and fine-level decoder network to produce coarse and fine tangent patch depth maps which are merged to get low and final high-resolution ERP depth maps respectively at the output. We evaluate our method using multiple real-world depth estimation benchmark datasets. We also perform an ablation study to analyze the proposed design choices. The conducted experiments show that our model outperforms the existing methods both quantitatively and qualitatively.

The following summarizes the main contributions of the paper:

- We present a 360 monocular depth prediction pipeline that addresses the existing depth map discrepancy issue via a novel coarse-fine feature fusion framework.
- We introduce a Multi-Scale Feature Fusion (MSFF) network to provide integrated contextual guidance to the local tangent patch features at the UNet-type network bottleneck, producing structure detailed, smooth, and consistent ERP depth maps.
- Our framework outperforms the existing methods on multiple real-world benchmark datasets both quantitatively and qualitatively.

2 RELATED WORK

2.1 Monocular perspective depth estimation

Initially, researchers solved scene understanding problems using a sequence of images from video or stereo image pair matching approach [17, 22, 41]. These approaches, however, are not feasible due to the dataset requirement and suffer degraded performance in textureless and repetitive regions. Therefore, with the advancement of deep learning technology, the research community started working on monocular perspective image depth estimation methods [3, 16, 24, 26, 40] that use variations of deep CNN architectures. For example, Laina et al. [24] uses a deep CNN with a residual learning framework. While methods in [3, 26, 40] use aggregation of CNN and fully connected conditional random field to combine multi-scale feature information for better scene understanding. However, the perspective images in these methods lack integrated scene representation due to their limited field of view generating inconsistent depth estimations.

2.2 Monocular 360 depth estimation

ERP images: With the increasing popularity of 360 photography, many researchers are motivated to use 360 images directly for depth estimation networks as they provide comprehensive scene representation with a wide field of view. However, commonly used ERP 360 images suffer from significant distortions in the top and bottom latitude regions. Therefore, using conventional neural networks developed for perspective images shows degraded performance for 360 images. To tackle this problem explicitly methods in [5, 8–11, 14, 15, 20, 32, 33, 36, 43–45] proposed using distortion aware or spherical kernel CNNs. For example, the methods in [5, 36] introduced distortion-aware deformable convolution to improve generalizability. ACDNet [44] showed that adaptive dilated convolution with channel-wise feature fusion produce better results than the deformable convolution approach. While methods in [8–10, 14, 15, 20, 32, 33, 43] proposed spherical CNNs to mitigate ERP distortion problem for depth estimation. These specifically designed spherical kernels and distortion-aware convolutional neural networks suffer from the performance limitations due to the utilization of fixed sampling positions.

Cube map and ERP images: Recently some methods instead of directly designing networks to handle ERP spherical distortion, use distortion-less 360 image representation input for performance improvement. For example, SliceNet [28] divides 360 images into vertical slices of the sphere and uses an encoder-decoder neural network. HoHoNet [35] learns compact latent horizontal 360 image features, however, it does not preserve boundaries and smoothness in the depth map. Hsien-Tzu Cheng et al. [6] uses distortion-less

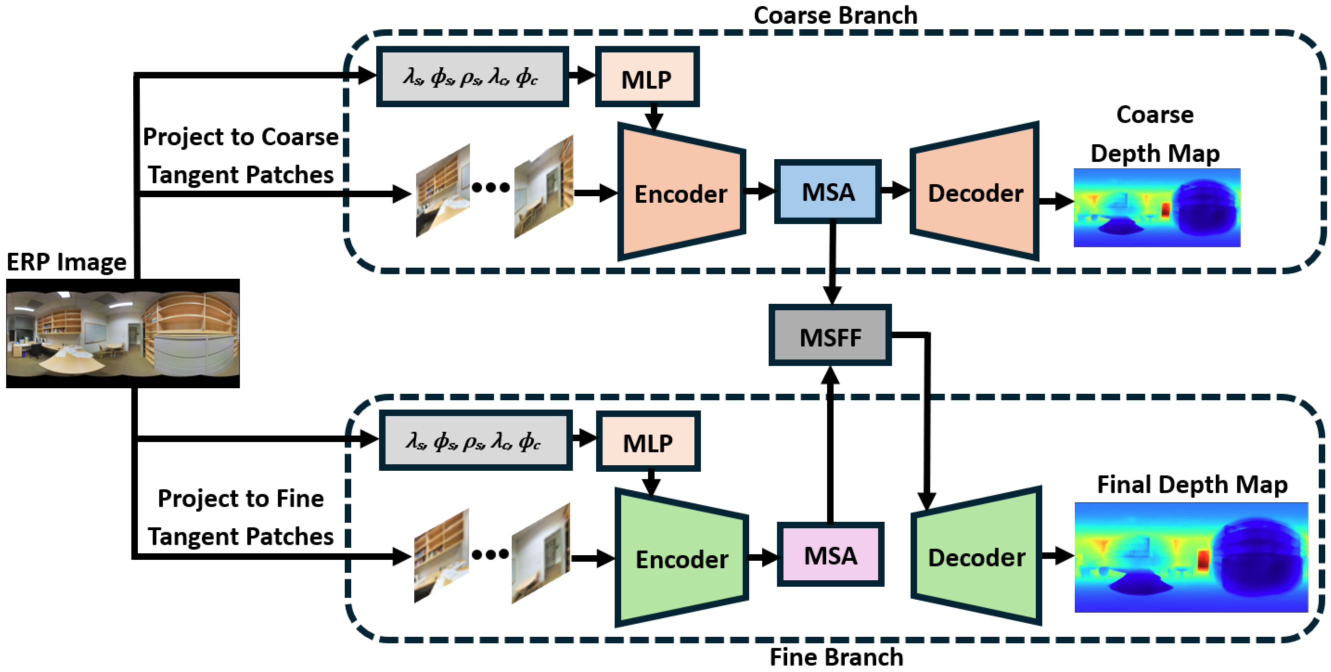

Figure 2: Overall pipeline of our method, MS360.

360 image projection called cube map as input to CNN, but these cube map patches input hinders the performance by the restricted field of view and face discontinuity. The methods in [21, 38, 39, 42] use a combination of ERP and cube map representation input, where method BiFuse [38] fuse ERP and cube map feature at the encoder-decoder network, UniFuse [21] and BiFuse++ [39] fuse these features only at the decoder stage and Zhiqiang Yan et al. [42] use multi-model mask approach with ERP, cube map, and sparse depth as input for depth prediction. These ERP and cube map combination methods utilize complex and computationally heavy cross-projection fusion process and have been shown to suffer from cube map discontinuities and geometric irregularities in the final depth estimation.

Perspective tangent patch images: Lately, models in [7, 13, 25, 30, 31, 34] propose using multiple distortion-less perspective tangent patches projected from ERP images, allowing direct use of CNNs to estimate perspective depths that are merged to produce final ERP depth maps. For example, State-Of-The-Art (SOTA) OmniFusion [25] uses this strategy to predict perspective depth maps from multiple projected tangent patch images and merges them to produce a final ERP depth map using a geometric aware CNN and vision transformer. Although these models show better performance than the previously discussed approaches, their ERP depths show local patch merging artifacts. Also, since these tangent patch image features miss learning holistic ERP image information they produce inconsistent, uneven, and less accurate depth maps.

2.3 Attention based networks

Transformer attention-based network by Ashish Vaswani et al. [37] have shown prominent success in natural language processing tasks and currently are being increasingly used to solve computer vision problems. These networks with global receptive fields can capture long-range relationships using sequential multi-head attention blocks. René Ranft et al. [29] uses a vision transformer network to estimate depth using a single RGB image. Also, the latest method CMT [18] uses a combination of CNN and a transformer to capture local and global information to further improve the feature aggregation capability of the network. In our method, we use a multi-head self-attention (MSA) network by Ashish Vaswani et al. [37] at the bottleneck of the coarse and fine encoder-decoder Unet-type network to learn long-range context information amongst the multiple perspective distortion-free tangent patches.

3 METHODOLOGY

In this section, we discuss the details of our approach. The complete pipeline of our method is shown in Figure 2. Our framework focuses on addressing the discrepancies in the existing methods' depth maps due to the considered local tangent patch image feature missing the integrated image information. For this, we present a combination of coarse and fine branch networks that produce low and high-resolution ERP depth maps. The coarse branch takes a down-sampled ERP image input. This low-resolution ERP image is projected to multiple coarse tangent patch images, which are then provided to the encoder-decoder UNet network to produce coarse-level perspective depth maps. Our fine branch on the other hand utilizes the fine tangent patch images projected from high-resolution

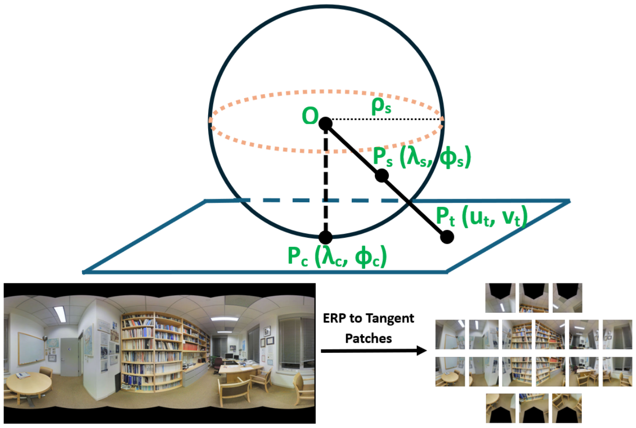

Figure 3: (Top) Diagram showing gnomonic projection [12] of point $P_s(\lambda_s, \phi_s)$ on a spherical surface onto point $P_t(u_t, v_t)$ on a tangent plane, where P_c is the center of the tangent plane whose spherical coordinates are denoted as (λ_c, ϕ_c) and ρ_s is the radius of the sphere (depth value in our case). (Bottom) Projection of tangent patches (with field of view = 80°) from ERP image.

ERP images. These patch images are again given to the UNet network to produce fine-level perspective depth maps. These predicted perspective depth maps from both branches are merged and projected back to get coarse and fine-level ERP depth maps. To improve the aggregation of local tangent patch image features we explicitly incorporate 3D geometric embedding information to the encoder network using an MLP network that takes tangent pixel spherical coordinates- $(\lambda_s, \phi_s, \rho_s)$ and patch center coordinates- (λ_c, ϕ_c) as input. Furthermore, to capture long-range global relationships among the local tangent patch image features Multi-head Self-Attention (MSA) [37] network is used at the bottleneck. Finally, to provide the fine-level image features with the guidance of coarse-level image features carrying comprehensive information we present a Multi-Scale Feature Fusion (MSFF) attention-based module. This module fuses coarse and fine-level bottleneck image features producing hybrid features, which are then provided to the fine branch decoder network to produce globally consistent, smooth, and more accurate depth maps.

3.1 Coarse and fine feature extraction

Our coarse branch extracts coarse-level image features. For this, we first down-sample the ERP image by factor 2. From the ERP image, we sample multiple (e.g. 18) tangent perspective images at different spherical latitudes (e.g: at ϕ_c = -67.5°, -22.5°, 22.5°, 67.5° latitudes with 3 (at λ_c = -120°, 0°, 120° longitude), 6 (at λ_c = -150°, -90°, -30°, 30°, 90°, 150° longitude), 6, 3 number of patches on each latitude respectively) using gnomonic projection [12] as shown in Figure 3. Equation 1 is used to convert the ERP image pixels (u_e, v_e) to the spherical coordinates $P_s(\lambda_s, \phi_s)$ first. Then using Equation 2 we obtain the projected tangent point $P_t(u_t, v_t)$ from spherical point $P_s(\lambda_s, \phi_s)$. We use Equation 3 to obtain tangent plane to spherical surface inverse gnomonic projection.

$$\lambda_s = 2\pi u_e/w,\ \phi_s = \pi v_e/h \tag{1}$$

where h and w are the height and width of the ERP image respectively.

$$\begin{aligned} u_t &= \frac{cos(\phi_s)sin(\lambda_s - \lambda_c)}{cos(\eta)} \\ v_t &= \frac{cos(\phi_c)sin(\phi_s) - sin(\phi_c)cos(\phi_s)cos(\lambda_s - \lambda_c)}{cos(\eta)} \\ cos(\eta) &= sin(\phi_c)sin(\phi_s) + cos(\phi_c)cos(\phi_s)cos(\lambda_s - \lambda_c) \end{aligned} \tag{2}$$

$$\begin{aligned} \lambda_s &= \lambda_c + tan^{-1}\Big(\frac{u_t sin(\eta)}{\gamma cos(\phi_c)cos(\eta) - v_t sin(\phi_c)sin(\eta)}\Big) \\ \phi_s &= sin^{-1}(cos(\eta)sin(\phi_c) + \frac{1}{\gamma} v_t sin(\eta)cos(\phi_c)) \end{aligned} \tag{3}$$

where, $\gamma = \sqrt{u_t{}^2 + v_t{}^2}$ and $\eta = tan^{-1}(\gamma)$. This way we obtain a one-on-one mapping between pixels of the tangent patch and the ERP image. The projected distortion-free multiple perspective tangent patches are then simultaneously given to the encoder network for feature extraction. The encoder network is initialized with ResNet [19] weights and consists of four stages of upsampling, 3x3 kernel-size convolution layers along with batch normalization, and a ReLU activation layer. Along with the tangent patches, their geometric information like in OmniFusion [25] is also provided to the encoder for better-aggregated feature extraction. For this, we use an MLP network that takes the spherical geometric position $(\lambda_s, \phi_s, \rho s)$ of a tangent pixel on the unit sphere and tangent patch center position (λ_c, ϕ_c) to generate geometric embedding of size $f_h \times f_w \times 64 \times N$ (where, f_h is feature map height, f_w is feature map width and N = number of tangent patches). The MLP network consists of two layers of convolution followed by batch-normalization and a ReLU activation layer. These geometric embedded feature maps are then concatenated with the first layer image features of the encoder for feature extraction. Additionally, we use an MSA network by Ashish Vaswani et al. [37] at the bottleneck to capture the global relationship amongst the local tangent patch image features. The input to the MSA module is the flattened encoded feature maps. These features are then used as tokens following the standard MSA transformer block architecture where the MSA output is computed using Equation 4 given below:

$$\begin{aligned} MSA(X) &= concat_{h=1}^{H}[Attn_h(X)] * W, \\ Attn_h(X) &= softmax(QK^T/\sqrt{d_h}) * V, \\ Q = X * W_Q&, K = X * W_K, V = X * W_V \end{aligned} \tag{4}$$

where Q, K, and V denote query, key, and value matrix, and W_Q, W_K, W_V represent their attention weights respectively, while h corresponds to the number of heads. We also reshape the dimensions of the MSA output to match the encoded feature map and integrate both for the decoder module input.

In contrast to the coarse, the fine branch uses high-resolution ERP images to produce fine-level tangent patches. Again the encoder and MSA network by Ashish Vaswani et al. [37] is utilized to produce the fine-level bottleneck image features.

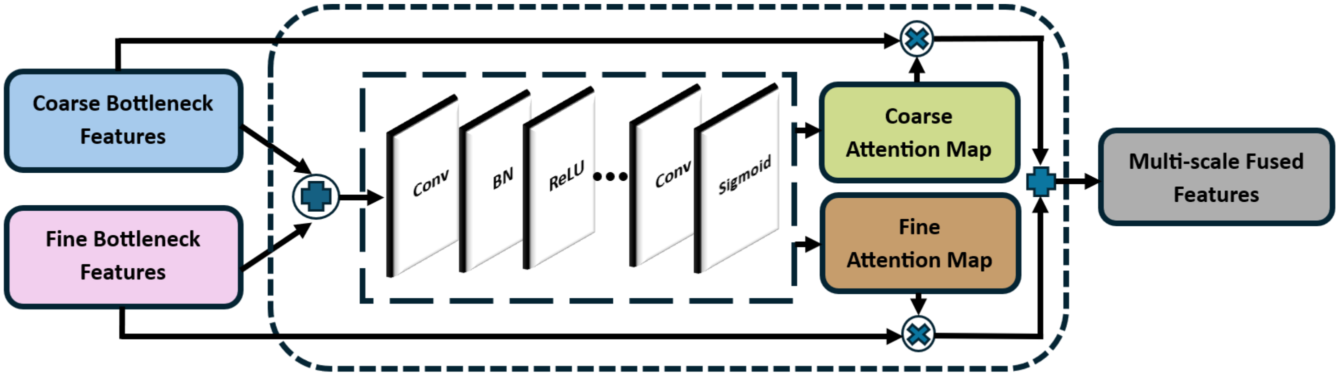

Figure 4: Overview of Multi-Scale Feature Fusion (MSFF) module, where Conv and BN stand for Convolution and Batch Normalization layer respectively.

3.2 Multi-scale feature fusion

The tangent patch image features only consider the local contextual information that leads to inconsistent ERP depth prediction resulting from merging local tangent depth predictions. To tackle this, we guide the fine-level image features with coarse-level integrated information by fusing coarse and fine-level features using an attention network. Inspired by the global to local feature fusion in GLPanoDepth [2], we combine the obtained multi-scale tangent image feature maps at the bottleneck using the Multi-Scale Feature Fusion (MSFF) module shown in Figure 4 which experimentally achieves the best performance. This module takes coarse and fine branch bottleneck encoded image features of size $f_{\text{th}} \times f_{\text{tw}} \times 512 \times N$ (where, f_{th} = tangent patch feature map height and f_{tw} = tangent patch feature map width) as input, which are then channel-wise concatenated and given to the three stages of convolution layer along with batch-normalization and ReLU activation layer followed by sigmoid activate layer to learn attention maps for coarse and fine image features. The coarse attention map focuses on capturing high-level comprehensive features while the fine attention map learns local information. These attention maps are multiplied to their respective feature maps and then aggregated to produce updated hybrid image features ($f_{\text{th}} \times f_{\text{tw}} \times 512 \times N$) carrying necessary global and local geometric information.

3.3 Coarse and fine depth estimation

The coarse bottleneck encoded features are given to the coarse branch decoder network to produce perspective depth maps which are merged back to low-resolution ERP format. To produce a high-resolution consistent ERP depth map the fine branch takes updated hybrid features from the MSFF module and gives them to the fine branch decoder network. Both the coarse and fine-level decoder networks consist of four stages of up-sampling, 3 x 3 kernel size convolution layers with batch-normalization, ReLU-activation layer, and skip connections from the corresponding four-stage encoder layers. To further refine the estimated ERP depth maps we use an iterative approach similar to OmniFusion [25] that uses the predicted depth values of an iteration to update the 3D geometric information ρ_s given to the MLP network for the next iteration.

3.4 Loss function

We supervise our network in an end-to-end manner using coarse and fine depth estimation loss denoted as L_{coarse} and L_{fine} respectively as shown in Equation 5. We use BerHu loss [24] to optimize coarse and fine depth estimation. The Equation of the BerHu depth loss function is given below:

$$\mathcal{L}_{coarse/fine} = \begin{cases} |\Delta D|, |\Delta D| \leq c \\ \frac{\Delta D^2 + c^2}{2c}, |\Delta D| > c \end{cases} \tag{5}$$

where, $|\Delta D| = |D_{gt} - D_{est}| * M$ is the L1 norm between the ground truth D_{gt} and the estimated D_{est} depth map. M is the valid depth pixel mask. And $c = 0.2max(\Delta D)$ is the 20% of the maximum per batch residual.

The total loss is defined as the addition of coarse and fine depth loss summed over all the iterations as shown in Equation 6:

$$\mathcal{L}_{total} = \sum_{iterations} L_{coarse} + \sum_{iterations} L_{fine} \tag{6}$$

4 EXPERIMENTS

4.1 Dataset and metrics

We perform the experiments on the widely known benchmark datasets called Stanford2D3D [1] and Matterport3D [4]. Stanford2D3D consists of large-scale real-world indoor scenes. In total, the dataset consists of 1413 panorama images out of which we use 1040 for training and 373 for testing according to the official train-test split. The Matterport3D consists of 10800 RGBD images. We again follow the official split that uses 8786 images for training and the rest for testing.

To evaluate the depth estimation performance of the network we utilize the widely used depth estimation metrics, mentioned in previous literature work [21, 24, 38], called Absolute Relative Error (Abs Rel), Squared Relative Error (Sq Rel), Root Mean Squared Error (RMSE), Root Mean Squared Error in logarithmic space (RMSE (log)) and threshold-based accuracy δ_t, where $t \in 1.25, 1.25^2, 1.25^3$.

Table 1: Quantitative depth estimation comparisons on Stanford2D3D [1] benchmark dataset. Our method outperforms the listed existing works for all the metrics, achieving state-of-the-art performance.∗ denotes we re-trained the method following the officially released code.

Methods	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE (log)↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
FCRN [24]	0.1837	-	0.5774	-	0.7230	0.9207	0.9731
RectNet [45]	0.1996	-	0.6152	-	0.6877	0.8891	0.9578
BiFuse [38]	0.1209	-	0.4142	-	0.8660	0.9580	0.9860
UniFuse [21]	0.1114	-	0.3691	-	0.8711	0.9664	0.9882
HoHoNet [35]	0.1014	-	0.3834	-	0.9054	0.9693	0.988
Panoformer[31]	0.1131	-	0.3557	-	0.8808	0.9623	0.9855
BiFuse++ [39]	0.1117	-	0.3720	-	0.8783	0.9649	0.9884
OmniFusion[25]∗	0.0943	0.0547	0.3582	0.1656	0.8999	0.9742	0.9914
Ours	**0.0899**	**0.0511**	**0.3317**	**0.1543**	**0.9152**	**0.9806**	**0.9925**

Table 2: Quantitative depth estimation comparisons on Matterport3D [4] benchmark dataset. Our model again outperforms the existing models for almost all the metrics. ∗ denotes we re-trained the method following the officially released code. Note: Panoformer uses 1296 testing samples vs 2014 samples used by all the other methods.

Methods	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE (log)↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
FCRN [24]	0.2409	-	0.6704	-	0.7703	0.9174	0.9617
RectNet [45]	0.2901	-	0.7643	-	0.6830	0.8794	0.9429
BiFuse [38]	0.2048	-	0.6259	-	0.8452	0.9319	0.9632
UniFuse [21]	0.1063	-	0.4941	-	0.8897	0.9623	0.9831
HoHoNet [35]	0.1488	-	0.5138	-	0.8786	0.9519	0.9771
Panoformer [31]	0.0904	-	**0.4470**	-	0.8816	0.9661	0.9878
BiFuse++ [39]	0.1424	-	0.5190	-	0.8790	0.9517	0.9772
OmniFusion [25]∗	0.0853	0.0616	0.5035	0.1472	0.9224	0.9791	0.9925
Ours	**0.0732**	**0.0612**	0.4982	**0.1430**	**0.9392**	**0.9808**	**0.9932**

4.2 Implementation details

Implementation of our network is done using Pytorch. A single Nvidia RTX 24 GB GPU is used for training the network. We initialize our encoder with the pre-trained ResNet [19] on ImageNet and use Adam optimizer [23] and the cosine annealing learning rate policy [27] with initial learning rate = 10^{-4}. We set the batch size = 2, patch size = 256 × 256, number of patches = 18, and patch field of view = 80°, and MSA network number of depth layers = 6 and number of heads = 4 for both coarse and fine branches. For our network input, we use an ERP image of 512 × 1024 resolution. Our network is trained in an end-to-end manner using BerHu loss [24] for coarse and fine predicted depth optimization. We train our network using the Stanford2D3D [1] and Matterport3D [4] datasets for 80 and 60 epochs respectively.

4.3 Performance comparison with SOTAs

Quantitative performance: Table 1. shows the quantitative performance comparison of our model with the related existing methods on the Stanford2D3D [1] dataset. As observed from the Table 1 our method outperforms all the existing methods for all the evaluating metrics. Compared to UniFuse [21] our model improves Abs Rel by 19.3% and RMSE by 10.1%. When comparing with HoHoNet [35] our model improves Abs Rel by 11.3% and RMSE by 13.5%. Compared with Panoformer [31] our model improves Abs Rel by 20.51% and RMSE by 6.7%. Our method outperforms the SOTA OmniFusion [25] by 4.7% Abs Rel and 7.4% RMSE. We also analyze the network performance on the larger Matterport3D [4] dataset as shown in Table 2. Here also our method outperforms all the listed existing methods for almost all the metrics.

Qualitative performance: In Figure 5. we qualitatively compare our model performance with one of the SOTA methods called OmniFusion [25] on the Stanford2D3D [1] dataset. We can observe that OmniFusion [25] predicted depth maps show local tangent patch merging artifacts highlighted by the red box. For example, the cross effect near the wall region in the third column image, and the horizontal blocky effect in the left region in the first column image. As discussed previously, in contrast to our method OmniFusion [25] learns the local tangent patch depth maps ignoring the comprehensive context information. Due to this, when the globally inconsistent overlapping tangent patch depth maps are merged, we see merging artifacts like a cross, horizontal lines, blocky effects, etc. However, our model output does not show these artifacts as our local tangent patch features are guided with coarse-scale context information using attention-based feature fusion which in turn helps to produce globally consistent and smooth depth predictions.

We also present the comparative qualitative results of our model with HoHoNet [35], UniFuse [21] and OmniFusion [25] on Stanford2D3D [1] and Matterport3D [4] datasets as shown in Figure 6 and 7 respectively. We observe that our model can recover more detailed structures, for example, bookshelves in the second and third column images of Figure 6. It can create sharper object boundaries/edges in the depth maps, for example, door and table boundaries in the first and second column images of Figure 6. Also, our model can preserve the global geometric structure continuity of the indoor scenery due to the proposed attention-based multi-scale feature fusion.

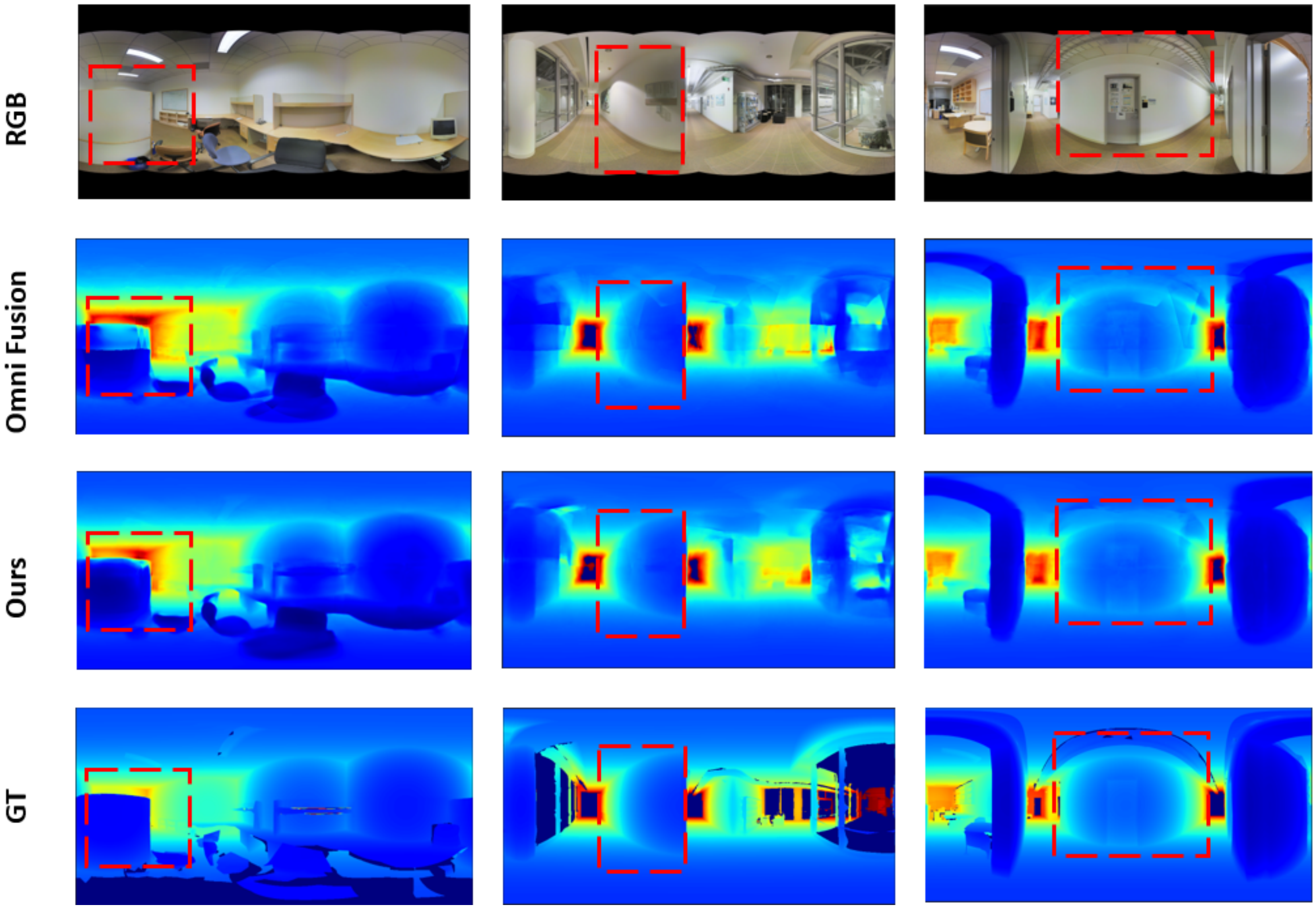

Figure 5: Qualitative comparisons of our method with one of the SOTA methods called OmniFusion [25] on Stanford2D3D [1] benchmark dataset. Our model-predicted depth maps do not show patch merging artifacts as seen in OmniFusion [25] (highlighted by red box). Our predicted depth maps are also consistent and have sharper object boundaries.

Table 3: Ablation study of the effect of MSFF module location, tangent patch Field Of View (FOV), tangent patch number, and coarse branch ERP image input resolution.

Parameter	Value	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE (log)↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
Multi-Scale Feature Fusion	Before MSA	0.0921	0.0548	0.3521	0.1653	0.9012	0.9746	0.9905
	After MSA	**0.0895**	**0.0520**	**0.3423**	**0.1577**	**0.9112**	**0.9759**	**0.9921**
FOV	120°	0.0925	0.0580	0.3550	0.1649	0.9098	0.9754	0.9909
	80°	**0.0895**	**0.0520**	**0.3423**	**0.1577**	**0.9112**	**0.9759**	**0.9921**
	40°	0.1286	0.0851	0.4443	0.2088	0.8336	0.9483	0.9824
No. of Patches	10	0.0965	0.0594	0.3541	0.1656	0.9061	0.9756	0.9901
	18	**0.0895**	**0.0520**	**0.3423**	**0.1577**	**0.9112**	**0.9759**	**0.9921**
Coarse ERP Input Resolution	128 × 256	0.0943	0.0553	0.3511	0.1655	0.9068	0.9742	0.9911
	256 × 512	**0.0895**	**0.0520**	**0.3423**	**0.1577**	**0.9112**	**0.9759**	**0.9921**

4.4 Ablation study

Study of MSFF module location: As observed from the quantitative comparisons, our proposed method achieves a performance boost by providing the missing comprehensive information to the local tangent patch features using coarse-fine feature fusion. In Table 3. we further analyze the effect of this MSFF module location in the framework. Adding an MSFF module before the MSA bottleneck module improves the performance over OmniFusion [25] while adding it after the MSA better helps to learn the necessary attention to guide the local bottleneck image feature with the coarse

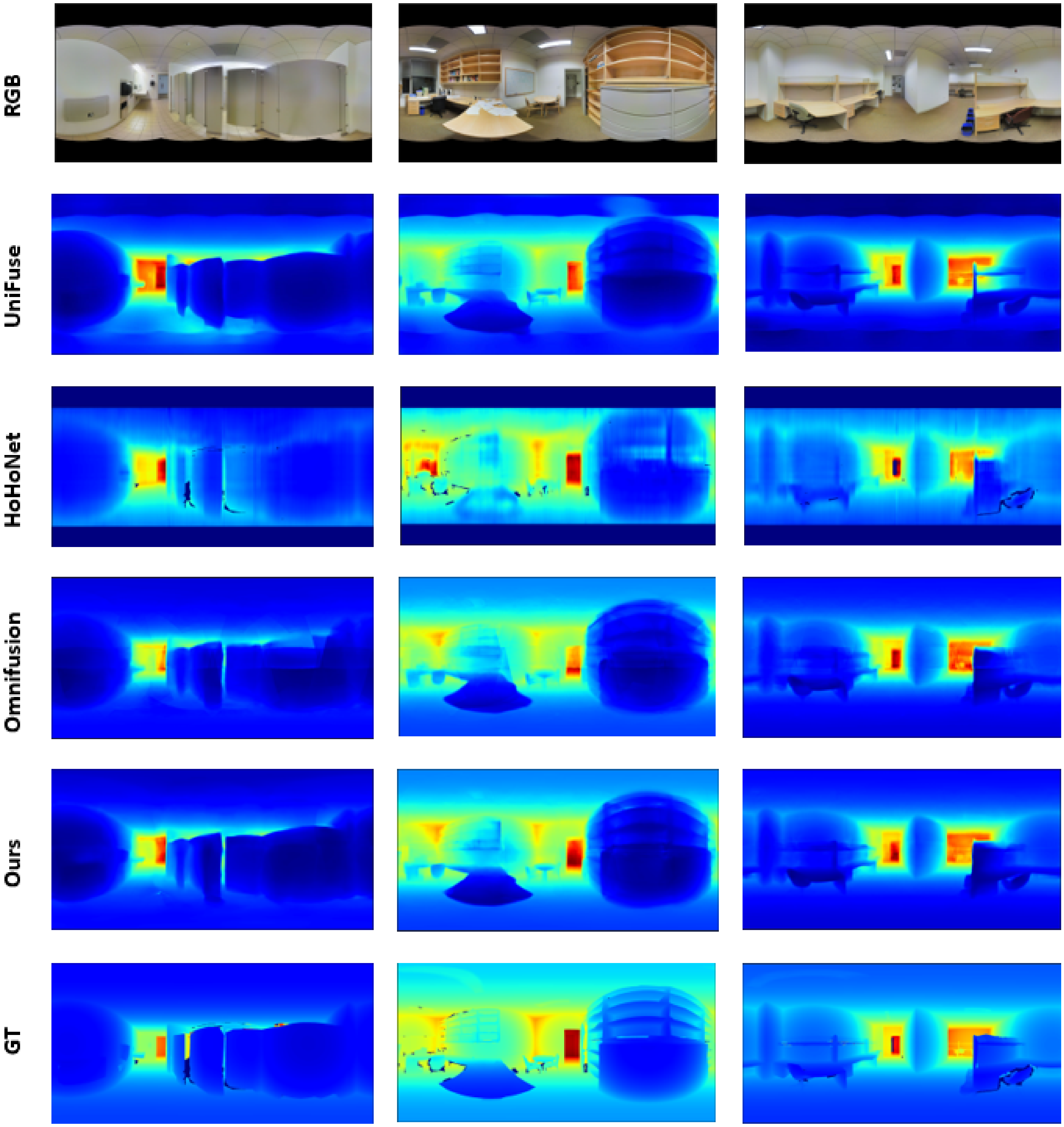

Figure 6: Comparative qualitative results on Stanford2D3D [1] benchmark dataset. We show the performance of UniFuse [21] (second row), HoHoNet [35] (third row), OmniFusion [25] (fourth row) and our model (fifth row) with RGB ERP input and Ground Truth (GT) depth map shown in first and last column respectively.

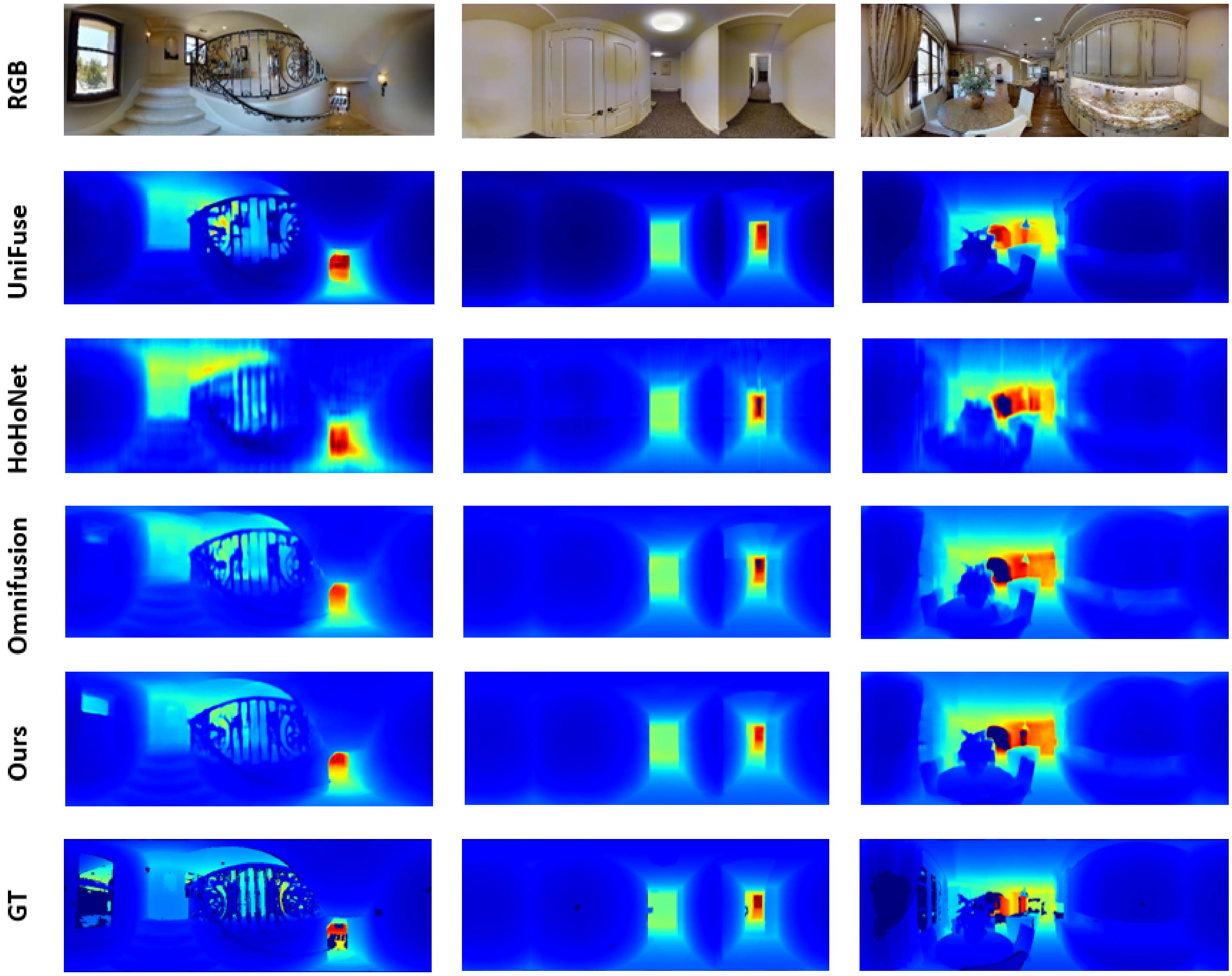

Figure 7: Comparative qualitative results on Matterport3D [4] benchmark dataset. We show the performance of UniFuse [21] (second row), HoHoNet [35] (third row), OmniFusion [25] (fourth row) and our model (fifth row). The RGB ERP input and Ground Truth (GT) depth map are shown in the first and last column respectively.

one giving the best performance. This is because after MSA the perspective tangent patch image features have better global relationship information which is further improved by our coarse-fine bottleneck image feature attention fusion.

Study of FOV, number of patches: Table 3. also shows the study of the effect of tangent patch Field Of View (FOV) and their number on the model performance. For the analysis of FOV, we fixed the number of patches to 18 and the patch size to 256×256 for coarse and fine branches. We observe that patch FOV= 80° achieves the best performance. When we decrease the patch FOV to 40° it degrades the performance as each tangent patch does not carry the required local information necessary to predict a consistent perspective depth map. Also, increasing the tangent patch FOV to 120° leads to performance degradation as wide FOV increases the overlapping region between the multiple tangent patches, leading to increased discrepancies among the predicted perspective depth maps.

Next, we study the effect of decreasing the number of tangent patches on the model performance. For this study, we fixed the patch FOV to 80° and the patch size to 256 × 256 again for both the coarse and fine branches. Ideally, we want less number of patches to make the model computationally efficient. We observe that decreasing the number of patches to 10 degrades the model performance as fewer patches do not cover enough regional information to predict the structure details in the depth maps.

Study of coarse branch ERP image input resolution: We also studied the effect of reducing the input ERP image resolution for the coarse branch as shown in Table 3. We observe that the model performance degrades when we reduce the image resolution to 128 × 256 from 256 × 512. This shows that very low-resolution input-extracted image features lack the necessary pattern details required for learning globally consistent depth estimation.

5 CONCLUSION

In this paper, we address the problem of discrepancy, discontinuity, and local patch merging artifacts present in the existing SOTA models' depth maps due to the missing integrated ERP input image feature learning. We present a novel two-branch coarse and fine network, that uses an UNet type encoder-decoder network with coarse and fine perspective tangent patches projected from ERP images as an input to estimate low and high-resolution ERP depth values. To guide the local tangent patch perspective image features with the coarse comprehensive contextual information we fuse coarse and fine image features at the network bottleneck using the attention-based Multi-Scale Feature Fusion (MSFF) module. Ablation studies were executed to analyze the proposed design choices to provide the best performance. Performed experiments show that our model produces geometric continuous, structurally detailed, accurate, and consistent depth maps outperforming the existing models on multiple monocular depth estimation benchmark datasets both quantitatively and qualitatively. In the future, we would like to extend our work by analyzing our model performance on the outdoor application datasets.

ACKNOWLEDGMENTS

This research was partially supported by the National Science Foundation under award CNS-2018850, the National Institute of Health under awards NIBIB-R01-EB02943, and U.S. Army Research Laboratory W911NF2120275. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the U. S. Government or agency thereof.

REFERENCES

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. 2017. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105* (2017).

[2] Jiayang Bai, Shuichang Lai, Haoyu Qin, Jie Guo, and Yanwen Guo. 2022. GLPanoDepth: Global-to-Local Panoramic Depth Estimation. arXiv:2202.02796 [cs.CV]

[3] Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. 2017. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology* 28, 11 (2017), 3174–3182.

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. *International Conference on 3D Vision (3DV)* (2017).

[5] Hong-Xiang Chen, Kunhong Li, Zhiheng Fu, Mengyi Liu, Zonghao Chen, and Yulan Guo. 2021. Distortion-aware monocular depth estimation for omnidirectional images. *IEEE Signal Processing Letters* 28 (2021), 334–338.

[6] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. 2018. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1420–1429.

[7] Shih-Han Chou, Yi-Chun Chen, Kuo-Hao Zeng, Hou-Ning Hu, Jianlong Fu, and Min Sun. 2018. Self-view grounding given a narrated 360 video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[8] Taco Cohen, Mario Geiger, Jonas Köhler, and Max Welling. 2017. Convolutional networks for spherical signals. *arXiv preprint arXiv:1709.04893* (2017).

[9] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. 2019. Gauge equivariant convolutional networks and the icosahedral CNN. In *International conference on Machine learning*. PMLR, 1321–1330.

[10] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. 2018. Spherical cnns. *arXiv preprint arXiv:1801.10130* (2018).

[11] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. 2018. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*. 518–533.

[12] H.S.M. Coxeter and H.S.M. Coxeter. 1989. *Introduction to Geometry*. Wiley. https://books.google.com/books?id=c0ld-crynsIC

[13] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. 2020. Tangent images for mitigating spherical distortion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12426–12434.

[14] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. 2018. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 52–68.

[15] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and Jose J Guerrero. 2020. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters* 5, 2 (2020), 1255–1262.

[16] 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.

[17] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2495–2504.

[18] Jianyuan Guo, Kai Han, Han Wu, Yehui Tang, Xinghao Chen, Yunhe Wang, and Chang Xu. 2022. CMT: Convolutional Neural Networks Meet Vision Transformers. arXiv:2107.06263 [cs.CV]

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

[20] Chiyu Jiang, Jingwei Huang, Karthik Kashinath, Philip Marcus, Matthias Niessner, et al. 2019. Spherical CNNs on unstructured grids. *arXiv preprint arXiv:1901.02039* (2019).

[21] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. 2021. Unifuse: Unidirectional fusion for 360 panorama depth estimation. *IEEE Robotics and Automation Letters* 6, 2 (2021), 1519–1526.

[22] Kevin Karsch, Ce Liu, and Sing Bing Kang. 2014. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE transactions on pattern analysis and machine intelligence* 36, 11 (2014), 2144–2158.

[23] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[24] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. 2016. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*. IEEE, 239–248.

[25] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. 2022. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2801–2810.

[26] Fayao Liu, Chunhua Shen, and Guosheng Lin. 2015. Deep convolutional neural fields for depth estimation from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5162–5170.

[27] Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic Gradient Descent with Warm Restarts. arXiv:1608.03983 [cs.LG]

[28] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. 2021. Slicenet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11536–11545.

[29] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision Transformers for Dense Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 12179–12188.

[30] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 2022. 360monodepth: High-resolution 360deg monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3762–3772.

[31] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. 2022. PanoFormer: Panorama Transformer for Indoor 360 Depth Estimation. In *European Conference on Computer Vision*. Springer, 195–211.

[32] Yu-Chuan Su and Kristen Grauman. 2017. Learning spherical convolution for fast features from 360 imagery. *Advances in Neural Information Processing Systems* 30 (2017).

[33] Yu-Chuan Su and Kristen Grauman. 2019. Kernel transformer networks for compact spherical convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9442–9451.

[34] Yu-Chuan Su, Dinesh Jayaraman, and Kristen Grauman. 2016. Pano2vid: Automatic cinematography for watching 360 videos. In *Asian Conference on Computer Vision*. Springer, 154–171.

[35] Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2021. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF*

Conference on Computer Vision and Pattern Recognition. 2573–2582.

[36] Keisuke Tateno, Nassir Navab, and Federico Tombari. 2018. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 707–722.

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[38] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. 2020. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 462–471.

[39] Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 2023. BiFuse++: Self-Supervised and Efficient Bi-Projection Fusion for 360° Depth Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 5 (2023), 5448–5460. https://doi.org/10.1109/TPAMI.2022.3203516

[40] Dan Xu, Wei Wang, Hao Tang, Hong Liu, Nicu Sebe, and Elisa Ricci. 2018. Structured attention guided convolutional neural fields for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3917–3925.

[41] Gangwei Xu, Xianqi Wang, Xiaohuan Ding, and Xin Yang. 2023. Iterative Geometry Encoding Volume for Stereo Matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21919–21928.

[42] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. 2022. Multi-modal masked pre-training for monocular panoramic depth completion. In *European Conference on Computer Vision*. Springer, 378–395.

[43] Jiachen Yang, Tianlin Liu, Bin Jiang, Wen Lu, and Qinggang Meng. 2020. Panoramic video quality assessment based on non-local spherical CNN. *IEEE Transactions on Multimedia* 23 (2020), 797–809.

[44] Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. 2022. ACDNet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3653–3661.

[45] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. 2018. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 448–465.