MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image

Uzair Shah, Muhammad Tukur, Mahmood Alzubaidi
ICT Division, College of Science and Engineering, Hamad Bin Khalifa University
Doha (Qatar)

Giovanni Pintore, Enrico Gobbetti
Visual and Data-intensive Computing, CRS4, Italy
National Research Center in HPC, Big Data, and Quantum Computing, Italy
(giovanni.pintore|enrico.gobbetti)@crs4.it

Mowafa Househ, Jens Schneider, Marco Agus
ICT Division, College of Science and Engineering, Hamad Bin Khalifa University
Doha (Qatar)
(magus|jeschneider)@hbku.edu.qa

Abstract

We present a novel holistic deep-learning approach for multi-task learning from a single indoor panoramic image. Our framework, named MultiPanoWise*, extends vision transformers to jointly infer multiple pixel-wise signals, such as depth, normals, and semantic segmentation, as well as signals from intrinsic decomposition, such as reflectance and shading. Our solution leverages a specific architecture combining a transformer-based encoder-decoder with multiple heads, by introducing, in particular, a novel context adjustment approach, to enforce knowledge distillation between the various signals. Moreover, at training time we introduce a hybrid loss scalarization method based on an augmented Chebychev/hypervolume scheme. We illustrate the capabilities of the proposed architecture on public-domain synthetic and real-world datasets. We demonstrate performance improvements with respect to the most recent methods specifically designed for single tasks, like, for example, individual depth estimation or semantic segmentation. To our knowledge, this is the first architecture capable of achieving state-of-the-art performance on the joint extraction of heterogeneous signals from single indoor omnidirectional images.*

1. Introduction

Spherical cameras offer cost-effective and efficient means to swiftly capture the complete surroundings around

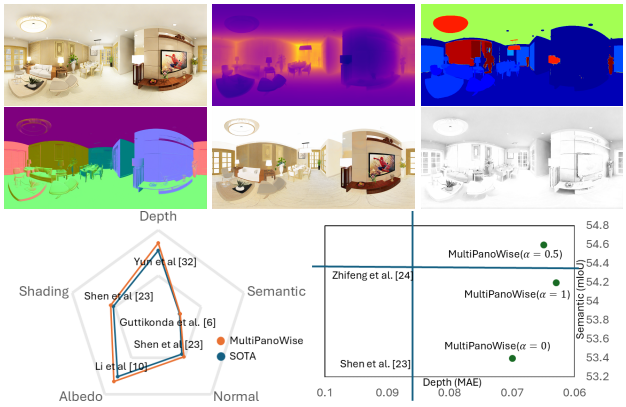

Figure 1. Our MultiPanoWise architecture can provide high accuracy joint dense predictions from single panoramic images. Top: example of multiple inferences obtained with MultiPanoWise on a single synthetic RGB from Structured3D [44]. From top to bottom, left to right, the RGB input, depth prediction, semantic inference, color-coded normals, reflectance, and shading. Bottom left: radar plot comparing the performance of joint prediction of MultiPanoWise with respect to the state-of-the-art single prediction for different signals on Structured3D dataset [44]. Bottom right: the performance obtained with different values of the hyperparameter α (see Sec. 3.2) compared to the state-of-the-art on real-world Stanford2D3D dataset [2]. For $\alpha = 0.5$ the model trained by MultiPanoWise can reach state-of-the-art performance for both semantic and depth prediction. For the accurate numerical values associated with these plots, we refer readers to Tab. 1,2 and 3.

the capture position in a single shot. The captured *panoramic* image, also called 360°, *omnidirectional*, or *surround-view* image, provides a wealth of information, and is especially suited to be exploited for designing or experiencing indoor environments. In particular, 360° views enable designers and developers to create more engaging and realistic environments, facilitate a deeper understanding of spatial layouts, enhance the visualization of potential design modifications, and let users experience virtual spaces more interactively and comprehensively [17, 21, 22].

Image-based design methods are gaining traction in both design and digital content creation, driven by their ability to deliver rich and immersive experiences. These methods, however, depend heavily on accurately estimating various types of information, such as depth signals for understanding spatial relationships, semantic segmentation for differentiating between object types, and material properties for realistic lighting effects and the seamless insertion of virtual objects. Accurate information estimation is crucial for creating believable and interactive digital environments, enhancing the realism and utility of virtual spaces for users.

Despite significant advances in image-based design technologies, a notable gap remains in multi-task inference systems tailored for indoor panoramic images. Previous methods have shown proficiency in extracting singular signals with remarkable accuracy [6, 26]. However, the complexity of panoramic imagery, particularly in indoor environments, demands a holistic approach that can concurrently process and interpret multiple types of information. This lack of integrated solutions presents a critical limitation, hindering the full potential of panoramic images in applications requiring comprehensive scene understanding and interaction.

To overcome such limitations, we introduce *MultiPanoWise*, a transformer-based holistic architecture for multi-signal inference from indoor panoramic images. Our approach is specifically targeted to exploit the unique characteristics of indoor 360° imagery and to the inherent consistency between various signals. To perform multi-task dense prediction, we introduce a transformer-based branched encoder-decoder architecture, where a common encoder-decoder network feeds multiple heads for dense estimation. The transformer-based encoder leverages the PanoFormer [26] baseline to generate features. The decoder progressively refines the encoded features through multiple convolutional layers boosted by a panoramic-specific self-attention mechanism [26] and feeds a set of convolution-based heads for dense prediction of the various signals. On top of this multi-head encoder-decoder architecture, we integrate the following main contributions:

- we introduce a context adjustment layer able to enforce knowledge distillation between the encoder-decoder and the various heads through skip connections stemming from the encoding layers (Sec. 3.1). In this way, the model can distill the relevant feature channels that can help in refining the low-resolution signals generated by the multiple heads;
- for the first time in the multi-task dense prediction domain, we introduce the usage of augmented hypervolume loss scalarization methods [42], that have proven to provide Pareto-optimal solutions in standard multi-task problems. To this end, we propose a hybrid Chebychev/linear scalarization scheme depending on a single parameter and able to pacify potentially conflicting prediction tasks during the training stage, without compromising the stability of gradients in the learning process (Sec. 3.2) in a way to boost prediction performance of multiple tasks concurrently (see Fig. 1 bottom).

Leveraging this architecture, MultiPanoWise can simultaneously process a single panoramic image to extract a comprehensive array of signals (see Fig. 1 top). These include geometric information represented as 16-bit depth and color-coded normals, dense semantic segmentation maps, and intrinsic decomposition signals distinguishing reflectance (albedo) and shading. This capability significantly improves processing efficiency and offers nuanced insights into indoor environments crucial for many applications, from virtual staging to advanced rendering techniques. We validated the proposed framework through comprehensive experiments on public domain real and synthetic indoor panoramic datasets such as Stanford2D3D [2], and Structured3D [44], where we obtained significant performance improvements over existing state-of-the-art methods tailored for individual signal inference. Finally, we showcase the positive effects of the context adjustment and the hybrid scalarization strategy through a dedicated ablation study.

2. Related work

Our framework deals with multi-task learning in the context of indoor panoramic images. For an extensive overview of the related work in these topics, we refer interested readers to the surveys about scene reconstruction from panoramic images [4, 23] and multi-task learning [43]. In the following, we discuss the methods that are most closely related to our work.

2.1. Inference from indoor omnidirectional images

Inferring geometric and physical signals from omnidirectional images is a challenging problem that attracted the computer vision community over the last few years. In general, to deal properly with the spherical distortion induced by equirectangular projection, various methods considering reprojection in combination with Convolutional Neural Net-

works have been considered. More specifically, various solutions have been proposed for individual dense estimation problems:

Depth estimation. The methods targeting depth estimation dramatically reduced the distortion by considering the conversion to sky-box representations. For example, UniFuse [8] and Bifuse [30] consider fusing features extracted from equirectangular projection and features extracted from cube maps at various stages of encoder-decoder architectures. At the same time, M3PT [33] uses random masking to process panoramas and sky-box depth patches simultaneously targeting panoramic depth completion. Other methods consider slicing the panoramic image along the vertical direction by assuming that vertical lines are not distorted by equirectangular projection and the acquisition is mostly gravity-aligned [19, 20]. Other methods consider tangent projection to extract multiple undistorted patches that can be processed through standard methods commonly used for perspective images, like 360MonoDepth [24] and OmniFusion [12]. Very recently, vision transformers have been exploited: PanoFormer [26] considers tangent projection (TP) to reduce the inherent distortion of omnidirectional images and uses the TP patches as tokens for a vision transformer architecture, while HRDFuse [1] propose a hybrid CNN-transformer architecture that integrates the holistic contextual information from the original equirectangular projection together with the regional structural information extracted from the tangent projection, and PanelNet [34] represents equirectangular projection (ERP) as consecutive vertical panels with corresponding panel geometry and uses it in a transformer for aggregating the local information within a panel with the panel-wise global context.

Semantic segmentation. Various architectures exploited the advantages of vision transformer and attention mechanisms to infer semantic segmentation: various methods consider extending high-performance segmentation transformer architectures to the spherical domain by incorporating geometric constraints, like Trans4Pass [38], or Trans4Pass+ [39], and by proposing distortion-aware attention (DA) attention schemes that capture the neighboring pixel distribution without using any geometric constraints [45]. Other methods instead consider multi-modal image modalities for improving segmentation, like the usage of depth signal, normal signal, or thermal signals: for example, AMBDRNet [40] consider fusing the thermal features with RGB images through bidirectional image-to-image translation, while CMX [37] considers cross-modal fusion with a transformer architecture in the context of perspective images, and the same concept is applied to indoor panoramic images by SFSS-MMSI [6] through the application of RGB, depth and normal signal in the context of Trans4Pass network. We consider a similar strategy for our architecture, but we exploit the concept of multi-modal fusion inherent in the multi-task learning of various signals, following the intuition of the existence of a significant correlation between the semantic signal and other signals, like normal and albedo, as shown by Baslamisli et al. [3].

Intrinsic image decomposition. Physically-based inverse rendering methods consider methods for extracting material and shading information. In general, these methods consider a pre-extraction technique to isolate specific illumination information in the form of spherical Gaussians, that can be used together with spherical warping to exploit this information in the context of relighting or rendering [13], or for extracting accurate reflectance information [31].

While all previous methods show accurate reconstruction of single signals, to our knowledge there is a lack of multi-task networks able to extract heterogeneous geometric, physical, and semantic information: we obtain this by leveraging a multi-decoder transformer-based architecture and multi-modal fusion during learning through context adjustment and hybrid loss scalarization in a way to exploit the correlation between the concurrently extracted signals.

2.2. Multi-task learning in panoramic images

Multi-task learning methods try to develop machine learning models that can perform concurrently multiple different tasks [28]. In various cases, they proved to offer advantages in terms of data efficiency, model convergence, and reduced model overfitting. In general multi-task setting presents many optimization challenges, making it difficult to realize large efficiency gains compared to learning tasks independently: to alleviate this issue, some methods propose loss weighting schemes [9], gradient projection strategies [35], or multi-objective optimization, with the overall objective of finding Pareto optimal solutions [25,41]. In this work we consider multi-task learning for dense prediction using standard encoder-decoder architectures: according to a recent taxonomy [28], the various methods can be subdivided into two main categories, based on where the interactions between the various tasks happen, i.e., locations in the architecture where information are exchanged or shared between tasks. Encoder-focused architectures share the task information in the encoder stage, before processing them with a set of independent task-specific decoders. In this category, Multi-Task Attention Networks (MTAN) [15] use a shared backbone network in conjunction with task-specific attention modules in the encoder and each task-specific attention module selects features from the general pool by applying a soft attention mask. Similarly, branched MTL networks [5] try to learn the hierarchical encoding structures inherent in images through ramifications starting with many shared layers, after which different tasks branch out into their sequence of layers. On the other side, some recent methods exchange information during the decoding

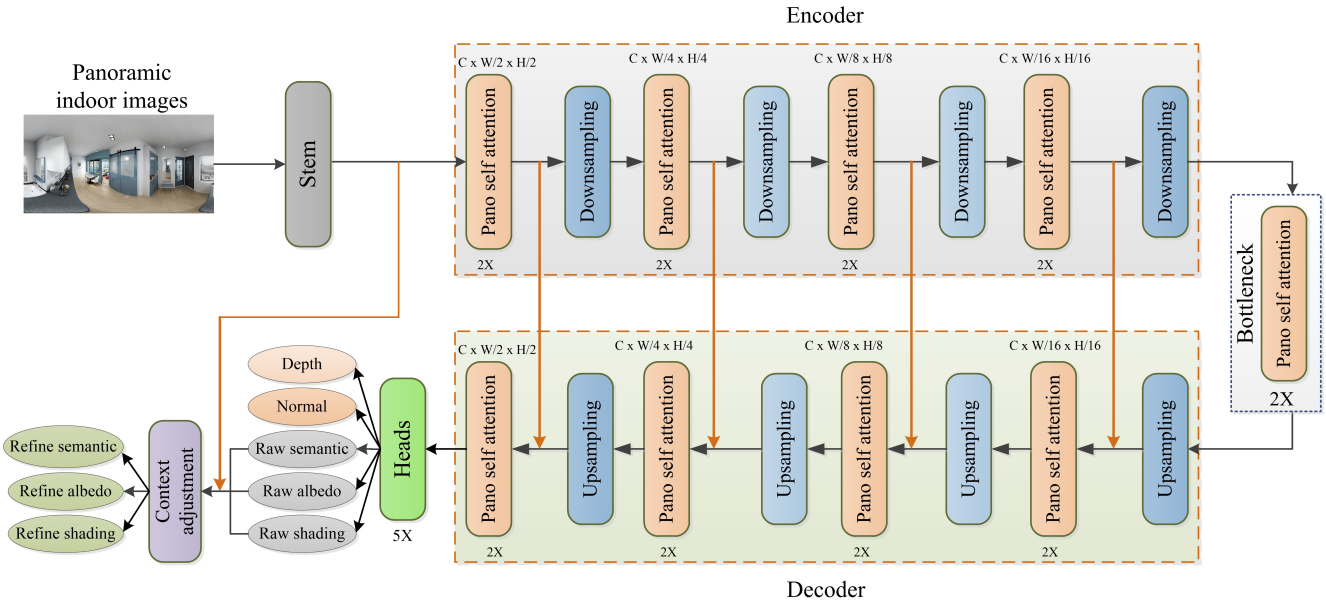

Figure 2. MultiPanoWise: The proposed architecture is built upon the PanoFormer architecture [26] We strictly adhere to their architectural design and various components in the encoder and decoder blocks. Additionally, we introduce a context adjustment layer to refine the raw semantic, albedo, and shading outputs. The raw semantic, albedo, and shading, along with the low-level features extracted from the stem, are fused and passed to the context adjustment layer to enhance the quality of semantic, albedo, and shading outputs.

stage: they normally employ a multi-task network to make initial task predictions, and then exploit the features from these initial predictions to refine each task output through knowledge distillation strategies [29,46]. Very recently, MtFormer [32] uses a vision transformer in which multiple tasks share the same transformer encoder and transformer decoder, and lightweight branches are introduced to harvest task-specific outputs, in a way to increase the MTL performance and reduce the time-space complexity. Similarly, Lopes et al. [16] consider pair-wise cross-task exchange through correlation-guided attention and self-attention for addressing multi-task learning in the context of indoor and outdoor semantic and geometric estimation from perspective images. In our method, we follow a similar strategy but we add a context adjustment layer based on skip connections to enforce signal refinement. For what concerns the multi-task training, we follow recent trends in loss scalarization [42], and we apply a hybrid linear/Chebychev scheme based on a single parameter. For what concerns the application of MTL to panoramic images, there is a lack of literature references: Li et al. [13] developed a multi-branch encoder-decoder architecture based on ResNet for inference of geometry and shading properties from single images, to perform inverse rendering, that we include in or comparison. Instead, we propose a transformer-based architecture, and we apply it for estimating heterogeneous signals, including the semantics. To the best of our knowledge, this is the first time vision transformers have been investigated to perform accurate multiple-dense predictions from single omnidirectional images.

3. Methods

We introduce MultiPanoWise, a multi-task learning model designed for panoramic images (an overview is shown in figure 2). We build our architecture upon the foundation laid by PanoFormer transformer [26], integrating a pixel-level patch division strategy for local feature enhancement, a relative position embedding method for improved positional information, and a panoramic self-attention mechanism to capture crucial panoramic structures essential for diverse signals. From the architectural perspective, we retain similar components starting from an input stem with a 3×3 convolution layer, followed by an encoder and a decoder, each comprising four hierarchical stages encompassing position embedding, two PST blocks, and a convolution layer. In addition, we augment the decoder with multiple heads to facilitate the generation of various signals. Moreover, we introduce a novel context adjustment module aimed at enhancing semantic segmentation and intrinsic decomposition signals such as reflectance and shading (described in section 3.1). The convolution layers within the encoder employ 4×4 kernels for dimension augmentation and down-sampling, while the decoder utilizes 2×2 transposed convolution layers for dimension reduction and up-sampling. Circular padding is utilized for all convolution layer padding operations.

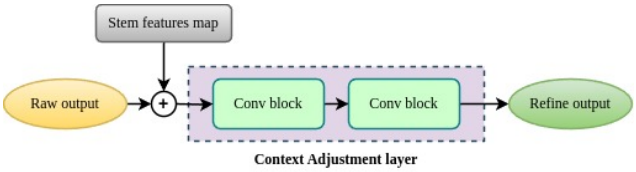

Figure 3. The detailed architecture of the context adjustment layer featuring two convolutional blocks.

3.1. Context adjustment layer

To improve the accuracy and reliability in extracting semantic, albedo, and shading information from the transformer model, we encounter a common challenge: the risk of data loss, often leading to unwanted gaps and distortions, especially at image edges. To address this challenge head-on, we introduce an additional component: the integration of a context adjustment layer as presented in figure 3. This layer leverages the rich low-level features extracted from the input stem layer, along with the raw semantic, albedo, and shading outputs from the joint decoder.

Through a fusion process, the feature maps extracted are seamlessly combined with the raw semantic maps from the joint decoder, resulting in composite images that blend high-resolution spatial details with semantic context seamlessly. This fusion process undergoes further refinement via a two-step convolutional block, consisting of a 3x3 convolutional layer followed by batch normalization and a Parametric Rectified Linear Unit (PReLU) activation function. This refinement step enhances the quality of information obtained from the raw semantic maps, yielding refined semantic representations. Similar procedures are applied to both reflectance and shading. This approach acts as a robust remedy for the inherent limitations of raw semantic, reflectance, and shading outputs. By effectively addressing distortions and artifacts, our network not only corrects imperfections but also preserves valuable data that might otherwise be lost. The resulting signals exhibit higher precision, characterized by sharper edges and enhanced fidelity (see Sec. 5).

3.2. Multi-task hybrid scalarization

Multi-task dense prediction can be considered a multi-objective optimization problem, in which the model needs to recover multiple signals together from a single image, and the different tasks are not necessarily aligned with each other, meaning that improving the prediction of one task might deteriorate the prediction of another task and vice versa. Recent efforts provided various methods for combining the multiple objectives trying to come out with theoretical solutions that are as optimal as possible [9, 25]. One of the typical solutions consists of scalarization [9], i.e. composing the various objective function in a single scalar function containing all tasks: given a vector of N tasks and the corresponding objective functions $\mathbf{L} = [\mathcal{L}_1, .., \mathcal{L}_N]$, scalar linearization involves a linear combination of the various losses:

$$\mathcal{L}_{lin} = \sum_{k=1}^{N} w_k \mathcal{L}_k, \tag{1}$$

in which the weights are learnable [9], while other strategies consider Chebychev/hypervolume scalarization

$$\mathcal{L}_{hyp} = \max_{k \in [1,N]} \{w_k \mathcal{L}_k\}, \tag{2}$$

that recently have been proven to converge to the Pareto front [41, 42]. Moreover, various groups considered augmenting hypervolume scalarization through regularization [14]: our method follows a similar strategy by defining a hybrid linear/hypervolume scalarization depending on a single hyperparameter α, and the scalar objective function is defined as follows:

$$\mathcal{L}_{hybrid} = \alpha \mathcal{L}_{hyp} + (1 - \alpha)\mathcal{L}_{lin}, \tag{3}$$

with $\alpha \in [0, 1]$, in a way that our loss scheme can range between a fully linear and fully Chebychev scalarization.

3.3. Task losses

For our multi-task dense prediction, we considered the following task losses:

- **depth, shading, normal, and reflectance estimation:** for all these tasks, we used the Berhu loss augmented with gradient loss based on Sobel filters detection [19, 26]

$$\mathcal{L}_x = \mathcal{B}_\beta(x, x_{gt}) + \mathcal{B}_\beta(S_H x, S_H x_{gt}) + \mathcal{B}_\beta(S_V x, S_V x_{gt}), \tag{4}$$

  where $x \in [d, sh, n, r]$,

$$\mathcal{B}_\beta(x, x_{gt}) = \begin{cases} \|x - x_{gt}\|, & \text{if } x < \beta \\ \frac{\|x - x_{gt}\|^2 + \beta^2}{2\beta}, & \text{if } x \geq \beta \end{cases}, \tag{5}$$

  and S_H, S_D are the classical Sobel gradient filters.

- **semantic segmentation**: we considered cross-entropy loss augmented with Dice loss

$$\mathcal{L}_s = \mathcal{C}(\mathbf{s}, \mathbf{s}_{gt}) + \mathcal{D}(\mathbf{s}, \mathbf{s}_{gt}), \tag{6}$$

  where

$$\mathcal{C}(\mathbf{s}, \mathbf{s}_{gt}) = -\log(\mathbf{s}_{gt}^T \mathbf{s}), \tag{7}$$

$$\mathcal{D}(\mathbf{s}, \mathbf{s}_{gt}) = \frac{\sum \mathbf{s}_{gt}^T \mathbf{s}}{2 \sum \mathbf{s}_{gt}}, \tag{8}$$

  and $\mathbf{s}$ is the one-hot encoding of the original label image s.

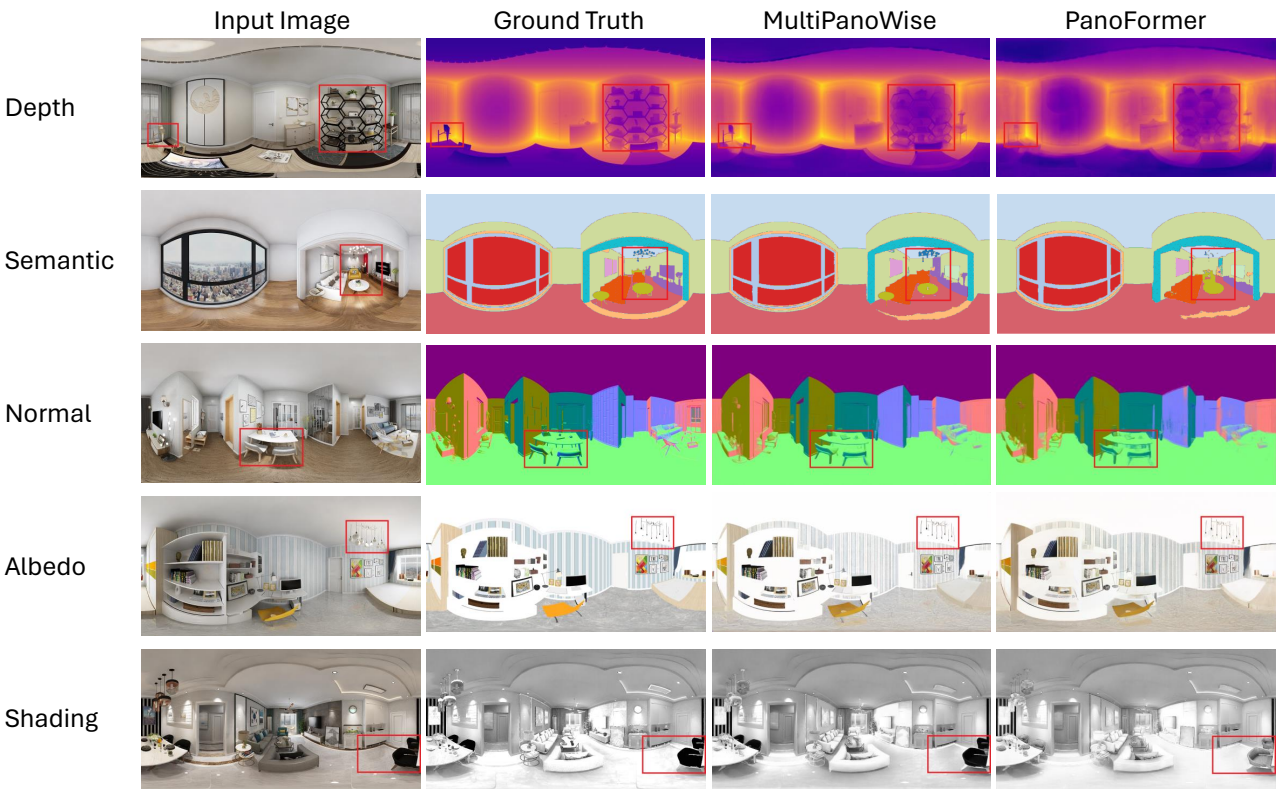

Figure 4. Qualitative comparison with Panoformer [26] trained for single inference on synthetic data (Structured3D). We showcase the inference for multiple signals (depth, albedo, normal, semantic and shading). For larger versions and zoomed insets, readers may check the supplementary material.

Table 1. Performance comparison on Structured3d dataset

Methods	Depth			Semantic	Albedo		Shading	Normal
	MAE ↓	MRE ↓	σ_1 ↑	mIoU ↑	MSE ↓	PSNR ↑	MSE ↓	MANE ↓
Li et. al [10]	n/a	n/a	n/a	n/a	0.073	n/a	n/a	24.7
Guttikonda et. al [6]	n/a	n/a	n/a	68.34	n/a	n/a	n/a	n/a
Yun et. al [36]	0.034	0.028	0.981	n/a	n/a	n/a	n/a	n/a
Pintore et. al [18]	0.091	0.054	0.954	n/a	n/a	n/a	n/a	n/a
Shen et al [26]	0.087	0.049	0.937	64.47	0.030	20.9	0.0916	7.25
MultiPanoWise (Ours)	0.056	0.019	0.975	69.61	0.021	22.45	0.0795	5.94

Table 2. Performance comparison on Stanford2d3d dataset.

Methods	Depth			Semantic
	MAE ↓	MRE ↓	σ_1 ↑	mIoU ↑
Ai et. al [1]	0.093	0.051	0.914	n/a
Li et. al [11]	0.165	0.084	0.929	n/a
Zhifeng et. al [27]	n/a	n/a	n/a	54.3
Guttikonda et. al [6]	n/a	n/a	n/a	52.87
Yu et. al [34]	0.152	0.083	0.924	46.3
Shen et al. [26]	0.086	0.050	0.934	51.2
MultiPanoWise (Ours)	0.065	0.038	0.945	54.6

4. Experiment setup

Datasets. In our experimentation, we assess the efficacy of our proposed approach using two indoor panoramic datasets: Structured3D (synthetic) and Stanford2D3D (real-world). The Structured3D dataset comprises 21,835 panoramic images, each meticulously annotated with detailed semantic, depth, normal, reflectance, and shading information. However, due to corrupted or incomplete annotations in some images, we meticulously cleaned the dataset, resulting in a final set of 17,434 images. Given the absence of an official dataset split, we adopt the split

outlined in [6] for model training and testing.

In contrast, the Stanford2D3D dataset contains 1,411 panoramic images, annotated with detailed semantic, normal, and depth information. For our experiments, we reserve area 5 for testing, while utilizing the remainder of the dataset for model training. Notably, we exclude the normal annotations from our consideration due to misalignment with the Structured3D dataset. Aligning these annotations necessitates extensive preprocessing and camera angle adjustments, making it impractical for our experimental setup. Therefore, we solely focus on the depth and semantic tasks for this dataset.

Implementation details. We implemented our method using PyTorch 2.1.2 and trained it on 8 Nvidia RTX 4090 GPUs. The Adam optimizer was employed with an initial learning rate of $1e^{-4}$, and the batch size was set to two. Images were resized to 512x1024 for both the training and testing phases. Initially, our model underwent training on the Structured3D dataset, chosen for its extensive size and diverse data representation. This training phase spanned 50 epochs, during which a multitask loss function was applied through simple addition scalarization. Subsequently, to refine the model's performance, we conducted fine-tuning using our hybrid scalarization scheme for an additional 20 epochs. Subsequently, we utilized the pre-trained model and fine-tuned it on the Stanford2D3D dataset for 20 epochs. During training, we applied various augmentations, including flipping, yaw rolling, and color adjustments such as brightness, contrast, and saturation set to 0.2, while hue was set to 0.1. It's important to note that augmentation was only applied to the Stanford2D3D dataset. While the weights and the hyperparameters are all learnable, this may cause strong gradient instabilities during the learning process. From our preliminary experiments, we realized that the range of the various objective functions was uniform, hence we fixed $w_k = 1$ for the remaining experiments, and the hyperparameter $\alpha = \frac{1}{2}$. Moreover, in Sec. 5.3 we report on an ablation study where we show how different values of α lead to different performance.

Evaluation Metrics. In our study, which involves multitasking, we have employed various evaluation metrics tailored to each task. For depth estimation, we have utilized three standard evaluation metrics commonly employed in the field: mean absolute error (MAE), mean square relative error (MRE), and the first threshold percentage, denoted as σ_1. In the context of semantic segmentation, we have calculated the Mean Intersection Over Union (mIoU). As for shading analysis, our evaluation includes the Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). When assessing albedo, we have considered the MAE, RMSE, and Peak Signal-to-Noise Ratio (PSNR). For color-coded normal surface estimation, the evaluation entails RMSE and Mean Angular Error (MAN).

5. Results

In this section, we provide a comprehensive evaluation of our Multi-Task model across various tasks in the 360D panoramic image, utilizing both synthetic and real-world datasets. We begin by comparing our results with recent studies focusing on single tasks. Furthermore, we delve into a detailed comparison with Panoformer [26], serving as a baseline, across multiple tasks to underscore the advantages of joint learning and context adjustment layers.

5.1. Structured3D

Depth. We employed metrics such as MAE, MRE, and $\sigma_{1.25}$ to assess the performance of our depth estimation. Table 1 showcases a quantitative comparison against the current state-of-the-art monocular panoramic depth estimation solutions. Our model outperforms the baseline across all metrics, demonstrating competitive performance compared to the current state-of-the-art. Particularly noteworthy is our achievement of 0.019, 0.056, and 0.975 for MRE, MAE, and $\sigma_{1.25}$, respectively. Moreover, our MultiPanoWise generates depth maps with enhanced accuracy, structure correlation, and semantic boundaries, as evidenced by the visual comparisons with Panoformer [26] in Figure 4.

Semantic. Similarly, we conducted a quantitative comparison of semantic segmentation results, as shown in Table 1. MultiPanoWise surpasses the current state-of-the-art by achieving a commendable mIoU of 69.61%. Notably, our model produces more accurate results compared to the baseline, as illustrated in Figure 4.

Normal. For normal estimation, we utilized Mean Angular Error as the validation metric. Our model achieved a MANE of 5.94, significantly outperforming previous work [10] and improving upon the baseline. Visual comparisons in Figure 4 highlight the superior accuracy of our MultiPanoWise in generating normal surfaces.

Albedo. Validation of albedo estimation was performed using MSE and PSNR. Our model demonstrates superior performance compared to both previous work and the baseline, achieving a PSNR of 22.45 and an MSE of 0.021, as indicated in Table 1. Visual inspection in Figure 4 further confirms the accuracy of our model in generating material colors.

Shading. While MSE was employed for performance measurement in shading estimation, no previous work was available for direct comparison. Our model achieved an MSE of 0.079, showcasing significantly improved color shading accuracy compared to the baseline's MSE of 0.916. Visual comparisons in Figure 4 underscore the superiority of our MultiPanoWise in generating accurate color shading.

5.2. Stanford2D3D

For the Stanford2D3D dataset, we focused on depth estimation and semantic segmentation: we excluded normals

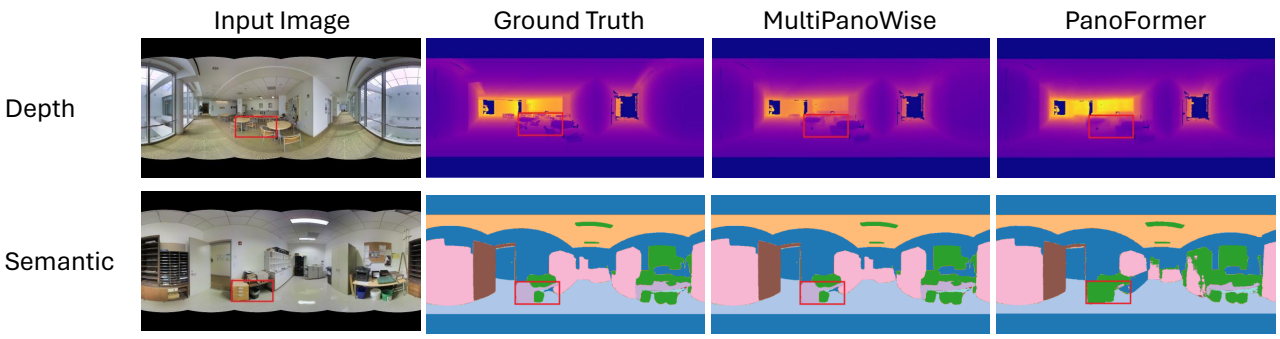

Figure 5. Qualitative comparison with Panoformer [26] trained for single inference on real data (Stanford2d3d). We showcase the inference for multiple signals (depth, semantic). For larger versions and zoomed insets, readers may check the supplementary material.

from this preliminary evaluation for camera orientation inconsistencies that require further processing.

Depth. Table 2 presents a quantitative comparison with the current state-of-the-art monocular panoramic depth estimation methods. Our model achieved impressive results with an MRE of 0.038, MAE of 0.065, and σ_1 of 0.945, placing first in the comparison.

Semantic. Table 2 outlines our model's performance in semantic segmentation, measured by mIoU. We attained a notable mIoU of 54.6, surpassing previous state-of-the-art methods. MultiPanoWise establishes a new best performance with a 54.6 mIoU, outperforming the single inference baseline, which achieved 51.2. We showcase a qualitative comparison with PanoFormer [26] trained for single inference in Fig. 5.

Table 3. Ablation on Stanford2d3d dataset.

Hybrid α	Fusion	Semantic(mIoU)	Depth(MRE)
0	✗	51.5	0.086
0.5	✗	52.9	0.072
1	✗	51.8	0.068
0	✓	53.4	0.068
0.5	✓	54.6	0.065
1	✓	54.2	0.063

5.3. Ablation study

In our ablation study, we maintained consistent hyperparameters as defaults for all experiments, following the settings outlined in Sec. 4. To assess the effectiveness of our model, we conducted experiments using the Stanford2D3D dataset. Given the proven effectiveness of the architectural components of PanoFormer [26], which we utilized in our model, our focus in the ablation study was on the context-adjustment layer and hybrid loss, controlled by parameter α (ranging from 0 for simple loss to 1 for full hypervolume loss). We began by modifying the Panoformer architecture to accommodate multitasking, achieved by adding multiple heads to shared decoders. Initially, we trained MultiPanoWise both with and without the context-adjustment layer on the Structured3D dataset, followed by fine-tuning on the Stanford2D3D dataset.

Table 3 summarizes our findings. Incorporating the context-adjustment layer results in an approximate 2.2% increase in mIoU, indicating improved semantic segmentation performance. Moreover, this addition also enhances depth estimation performance. Furthermore, employing the hybrid Chebychev/linear scalarization scheme yields over 1% improvement in mIoU, with a slight enhancement observed in the Mean Relative Error (MRE) of depth estimation.

6. Conclusion

We presented a novel holistic architecture for multi-task dense prediction on a single panoramic image representing indoor environments. We showcased that the proposed architecture can achieve state-of-the-art performance in typical dense prediction tasks of indoor panoramic images, like geometry estimation, semantic segmentation, or intrinsic image decomposition, by considering a context adjustment layer and a hybrid loss scalarization strategy. We believe that the proposed solution is general and might be applied to other multi-task dense prediction domains. We plan to extend it to more complex dense estimation problems, like signal extraction for inverse rendering [47] and virtual staging purposes [7].

Acknowledgments. This publication was made possible by NPRP-Standard (NPRP-S) 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). GP and EG also acknowledge the contribution of the Italian National Research Center in High Performance Computing, Big Data and Quantum Computing. The findings herein reflect the work and are solely the responsibility, of the authors.

References

[1] Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. Hrdfuse: Monocular 360deg depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13273–13282, June 2023. 3, 6

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

[3] Anil S Baslamisli, Thomas T Groenestege, Partha Das, Hoang-An Le, Sezer Karaoglu, and Theo Gevers. Joint learning of intrinsic images and semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 286–302, 2018. 3

[4] Thiago L. T. da Silveira, Paulo G. L. Pinto, Jeffri Murrugarra-Llerena, and Cláudio R. Jung. 3d scene geometry estimation from 360° imagery: A survey. *ACM Comput. Surv.*, 55(4), nov 2022. 2

[5] Pengsheng Guo, Chen-Yu Lee, and Daniel Ulbricht. Learning to branch for multi-task learning. In *International conference on machine learning*, pages 3854–3863. PMLR, 2020. 3

[6] Suresh Guttikonda and Jason Rambach. Single frame semantic segmentation using multi-modal spherical images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3222–3231, 2024. 2, 3, 6, 7

[7] Guanzhou Ji, Azadeh O Sawyer, and Srinivasa G Narasimhan. Virtual home staging: Inverse rendering and editing an indoor panorama under natural illumination. In *International Symposium on Visual Computing*, pages 329–342. Springer, 2023. 8

[8] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters*, 6(2):1519–1526, 2021. 3

[9] Alex Kendall, Yarin Gal, and Roberto Cipolla. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 3, 5

[10] Junxuan Li, Hongdong Li, and Yasuyuki Matsushita. Lighting, reflectance and geometry estimation from 360° panoramic stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 6, 7

[11] Meng Li, Senbo Wang, Weihao Yuan, Weichao Shen, Zhe Sheng, and Zilong Dong. $\mathcal{S}^2$net: Accurate panorama depth estimation on spherical surface. *IEEE Robotics and Automation Letters*, 8(2):1053–1060, 2023. 6

[12] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Duan Ye, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *2022 Conference on Computer Vision and Pattern Recognition (CVPR)*, New Orleans, USA, June 2022. 3

[13] Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, and Jiaqi Yang. Phyir: Physics-based inverse rendering for panoramic indoor images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12713–12723, June 2022. 3, 4

[14] Xi Lin, Zhiyuan Yang, Xiaoyuan Zhang, and Qingfu Zhang. Pareto set learning for expensive multi-objective optimization. *Advances in Neural Information Processing Systems*, 35:19231–19247, 2022. 5

[15] Shikun Liu, Edward Johns, and Andrew J. Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[16] Ivan Lopes, Tuan-Hung Vu, and Raoul de Charette. Cross-task attention mechanism for dense multi-task learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 2329–2338, January 2023. 4

[17] Ryan S Overbeck, Daniel Erickson, Daniel Evangelakos, Matt Pharr, and Paul Debevec. A system for acquiring, processing, and rendering panoramic light field stills for virtual reality. *ACM Transactions on Graphics (TOG)*, 37(6):1–15, 2018. 2

[18] Giovanni Pintore, Marco Agus, Eva Almansa, and Enrico Gobbetti. Instant automatic emptying of panoramic indoor scenes. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3629–3639, 11 2022. 6

[19] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. 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 (CVPR)*, pages 11536–11545, June 2021. 3, 5

[20] Giovanni Pintore, Eva Almansa, Marco Agus, and Enrico Gobbetti. Deep3DLayout: 3D reconstruction of an indoor layout from a spherical panoramic image. *ACM TOG*, 40(6):250:1–250:12, 2021. 3

[21] Giovanni Pintore, Eva Almansa, Armando Sanchez, Giorgio Vassena, and Enrico Gobbetti. Deep

panoramic depth prediction and completion for indoor scenes. *Computational Visual Media*, 2024. 2

[22] Giovanni Pintore, Fabio Ganovelli, Enrico Gobbetti, and Roberto Scopigno. Mobile mapping and visualization of indoor structures to simplify scene understanding and location awareness. In *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II*, pages 130–145. Springer, October 2016. 2

[23] Giovanni Pintore, Claudio Mura, Fabio Ganovelli, Lizeth Fuentes-Perez, Renato Pajarola, and Enrico Gobbetti. State-of-the-art in automatic 3D reconstruction of structured indoor environments. *Comput. Graph. Forum*, 39(2):667–699, 2020. 2

[24] Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 360MonoDepth: High-resolution 360deg monocular depth estimation. In *CVPR*, 2022. 3

[25] Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. *Advances in neural information processing systems*, 31, 2018. 3, 5

[26] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360 depth estimation. In *Computer Vision – ECCV 2022*, pages 195–211, Cham, 2022. Springer Nature. 2, 3, 4, 5, 6, 7, 8

[27] Zhifeng Teng, Jiaming Zhang, Kailun Yang, Kunyu Peng, Hao Shi, Simon Reiß, Ke Cao, and Rainer Stiefelhagen. 360bev: Panoramic semantic mapping for indoor bird's-eye view. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024. 6

[28] Simon Vandenhende, Stamatios Georgoulis, Wouter Van Gansbeke, Marc Proesmans, Dengxin Dai, and Luc Van Gool. Multi-task learning for dense prediction tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3614–3633, 2022. 3

[29] Simon Vandenhende, Stamatios Georgoulis, and Luc Van Gool. Mti-net: Multi-scale task interaction networks for multi-task learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 527–543. Springer, 2020. 4

[30] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[31] RongKai Xu, Lei Zhang, and Fanglue Zhang. Intrinsic omnidirectional image decomposition with illumination pre-extraction. *Authorea Preprints*, 2023. 3

[32] Xiaogang Xu, Hengshuang Zhao, Vibhav Vineet, Ser-Nam Lim, and Antonio Torralba. Mtformer: Multi-task learning via transformer and cross-task reasoning. In *European Conference on Computer Vision*, pages 304–321. Springer, 2022. 4

[33] Zhiqiang Yan, Xiang Li, Kun Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Multi-modal masked pre-training for monocular panoramic depth completion. In *European Conference on Computer Vision*, pages 378–395. Springer, 2022. 3

[34] H. Yu, L. He, B. Jian, W. Feng, and S. Liu. Panelnet: Understanding 360 indoor environment via panel representation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 878–887, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. 3, 6

[35] Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. *Advances in Neural Information Processing Systems*, 33:5824–5836, 2020. 3

[36] Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae Eun Rhee. EGFoRMeR: Equirectangular geometry-biased transformer for 360 depth estimation. *arXiv (Cornell University)*, 4 2023. 6

[37] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on Intelligent Transportation Systems*, 2023. 3

[38] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16917–16927, June 2022. 3

[39] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Philip HS Torr, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *arXiv preprint arXiv:2207.11860*, 2022. 3

[40] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2633–2642, June 2021. 3

[41] Richard Zhang and Daniel Golovin. Random hypervolume scalarizations for provable multi-objective black box optimization. In *International conference on machine learning*, pages 11096–11105. PMLR, 2020. 3, 5

[42] Xiaoyuan Zhang, Xi Lin, Bo Xue, Yifan Chen, and Qingfu Zhang. Hypervolume maximization: A geometric view of pareto set learning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 4, 5

[43] Yu Zhang and Qiang Yang. A survey on multi-task learning. *IEEE Transactions on Knowledge and Data Engineering*, 34(12):5586–5609, 2022. 2

[44] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer, 2020. 1, 2

[45] Xu Zheng, Tianbo Pan, Yunhao Luo, and Lin Wang. Look at the neighbor: Distortion-aware unsupervised domain adaptation for panoramic semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18687–18698, October 2023. 3

[46] Ling Zhou, Zhen Cui, Chunyan Xu, Zhenyu Zhang, Chaoqun Wang, Tong Zhang, and Jian Yang. Pattern-structure diffusion for multi-task learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4514–4523, 2020. 4

[47] Rui Zhu, Zhengqin Li, Janarbek Matai, Fatih Porikli, and Manmohan Chandraker. Irisformer: Dense vision transformers for single-image inverse rendering in indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2822–2831, June 2022. 8