MultiPanoWise: Holistic Deep Architecture for Multi-Task Dense Prediction from a Single Panoramic Image

Uzair Shah[1], Muhammad Tukur[1], Mahmood Alzubaidi[1], Giovanni Pintore[2], Enrico Gobbetti[3], Mowafa Househ[1], Jens Schneider[1], Marco Agus[1]
[1]College of Science and Engineering, Hamad Bin Khalifa University, Doha, Qatar
[2]Visual and Data-intensive Computing, CRS4, Italy
[3]National Research Center in HPC, Big Data, and Quantum Computing, Italy

جامعة حمد بن خليفة HAMAD BIN KHALIFA UNIVERSITY كلية العلوم والهندسة College of Science & Engineering
OmniCV2024
CVPR SEATTLE, WA JUNE 17-21, 2024

Introduction and Background

- **Objective**: Present a novel holistic deep-learning approach for multi-task learning from a single indoor panoramic image.
- **Background**: Explain the use of spherical cameras for capturing comprehensive indoor environments and the challenges of multi-task inference systems.
- **Key Points:**
  - Holistic approach for dense prediction.
  - *MultiPanoWise* leverages vision transformers.
  - Achieves state-of-the-art performance for joint extraction of multiple signals.
- **Importance of 360 Views**: Enhance understanding of spatial layouts, facilitate realistic environment visualization, improve design and user interaction.
- **Challenges and Proposed Solution**:
- Existing methods excel in single-task predictions but lack integrated multi-task capabilities.
- *MultiPanoWise* addresses this with a transformer-based architecture tailored for indoor 360° imagery.

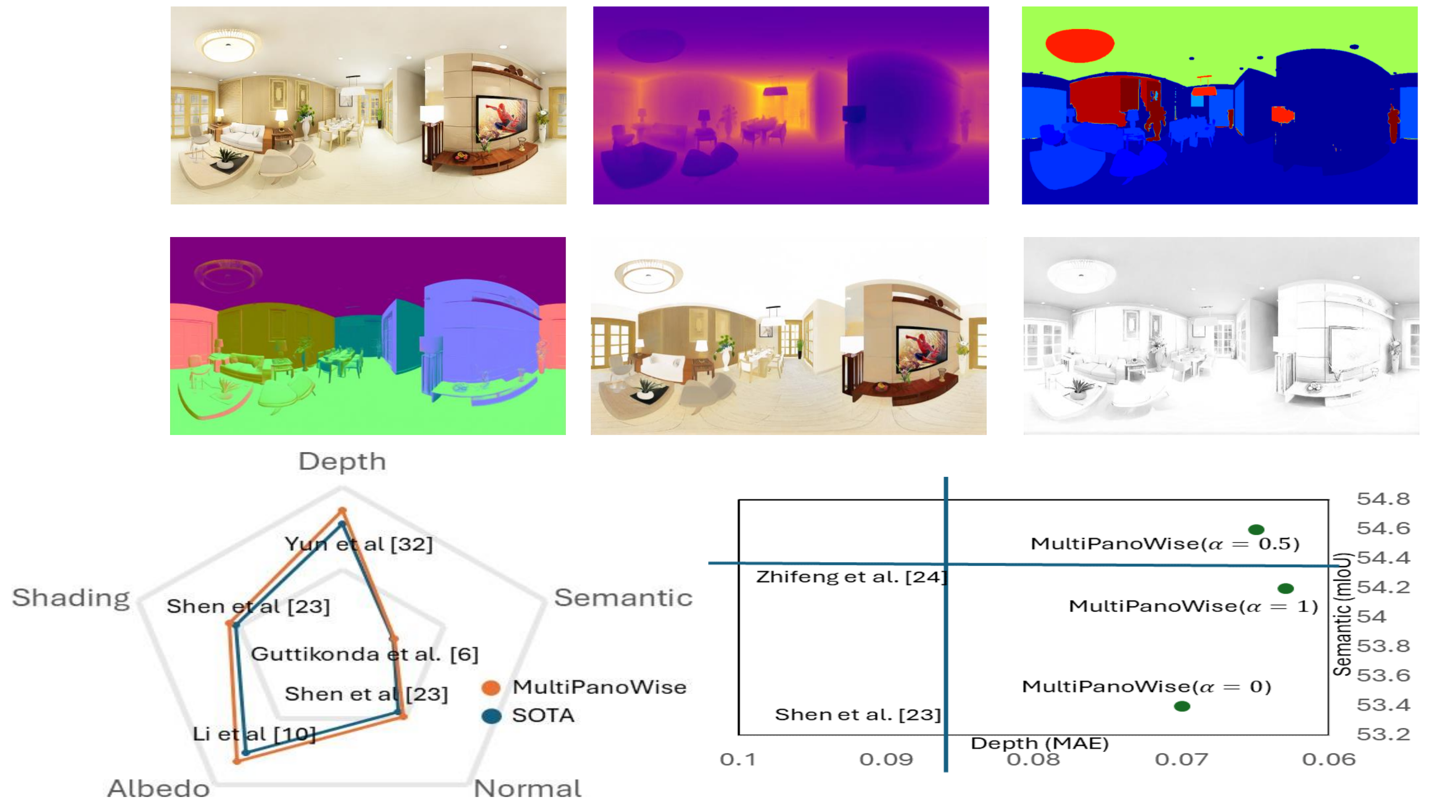


Figure 1. *MultiPanoWise* architecture provides accurate joint dense predictions from single panoramic images. Top: Example inferences on a synthetic RGB from Structured3D showing RGB input, depth, semantics, normals, reflectance, and shading. Bottom Left: Radar plot comparing *MultiPanoWise* with state-of-the-art single predictions on Structured3DBottom Right: Performance with different hyperparameter α values on Stanford2D3D. For α = 0.5, MultiPanoWise reaches state-of-the-art performance for both semantic and depth predictions.

Methods and Architecture

- **Architecture Overview:** *MultiPanoWise* extends PanoFormer with a multi-head encoder-decoder architecture for multi-task prediction.

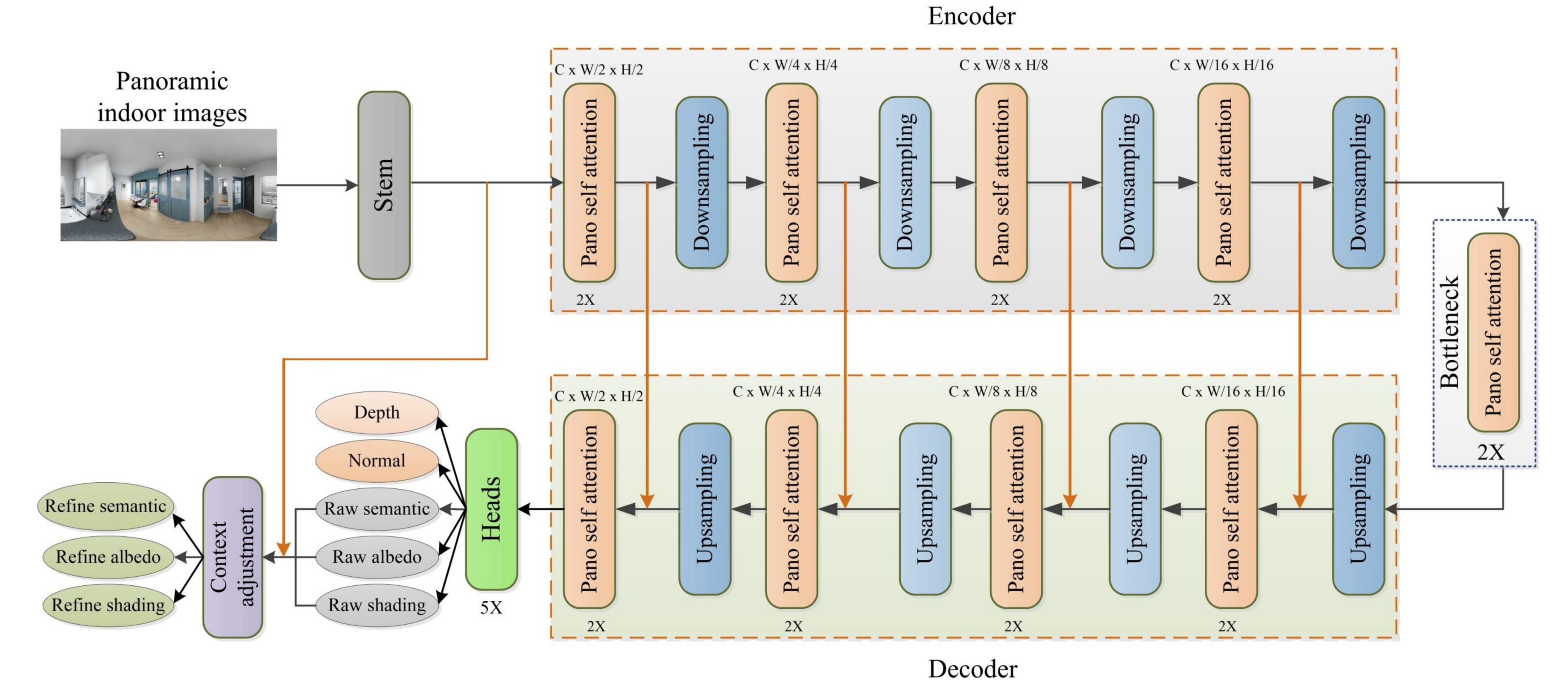


Figure 2. MultiPanoWise architecture, based on PanoFormer [1], includes a context adjustment layer to refine semantic, albedo, and shading outputs by fusing them with low-level features

- **Context Adjustment Layer**: Enforces knowledge distillation between the encoder and multiple heads and uses skip connections and refinement layers to enhance signal quality.

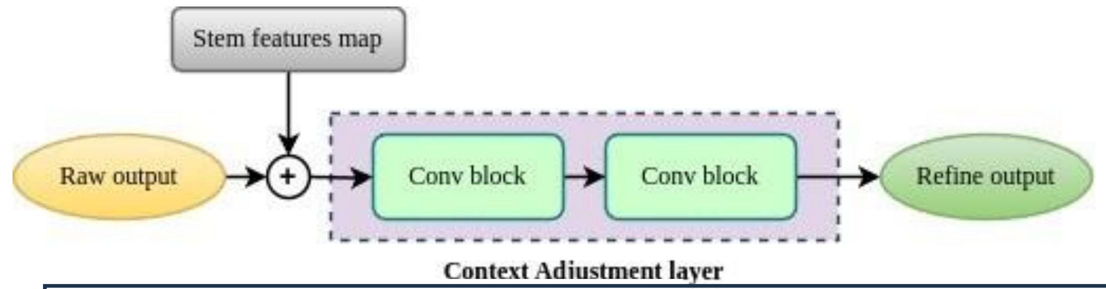


Figure 3. The detailed architecture of the context adjustment layer featuring two convolutional blocks.

$$\mathcal{L}_{lin} = \sum_{k=}^{N} w_k \mathcal{L}_k, \longrightarrow \mathcal{L}_{hyp} = \max_{k\in[1,N]} \{w_k \mathcal{L}_k\},$$

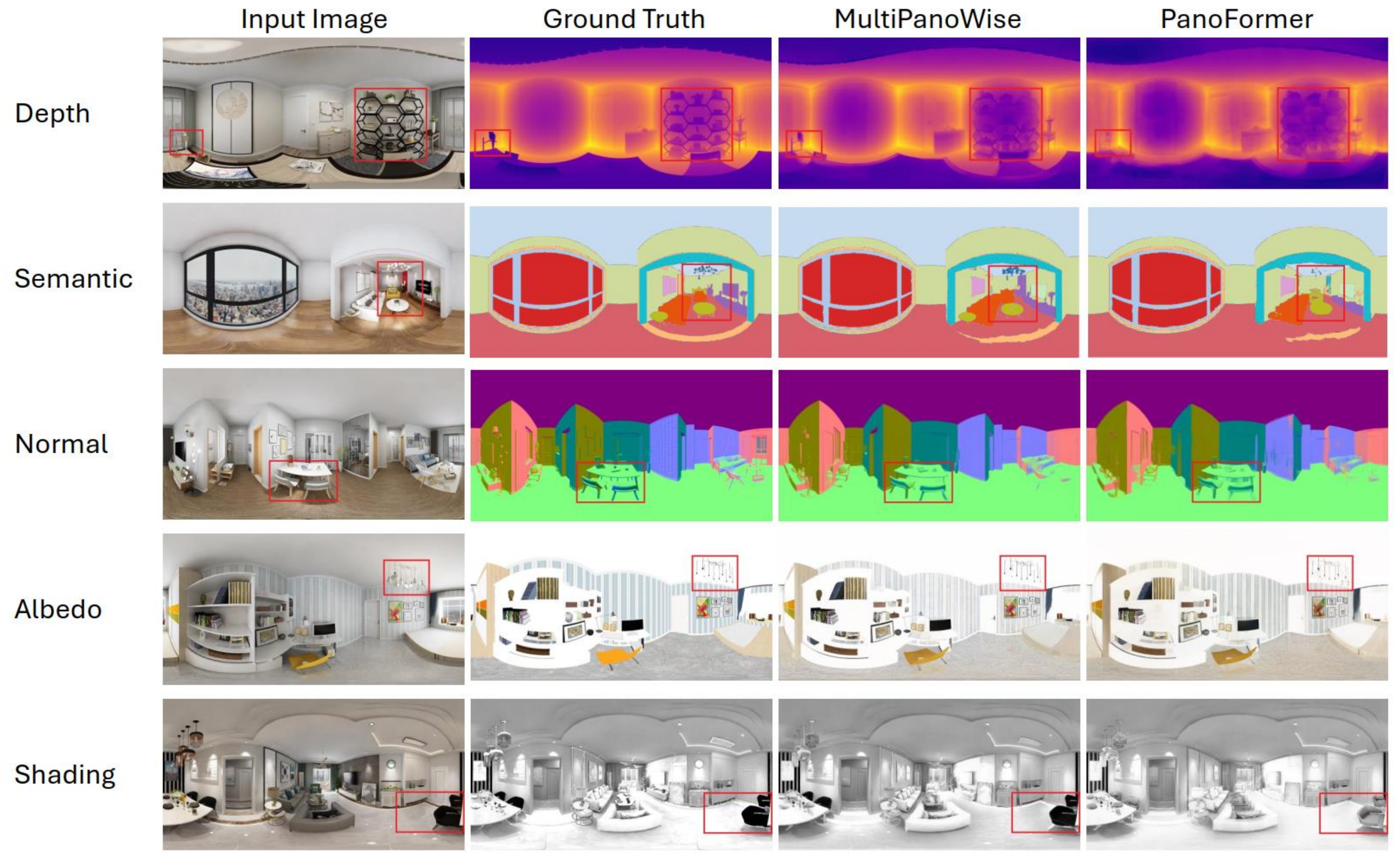
$$\mathcal{L}_{hybrid} = \alpha \mathcal{L}_{hyp} + (1-\alpha)\mathcal{L}_{lin},$$

- **Hybrid Loss Scalarization:** Balances the performance of multiple tasks during training and provides Pareto-optimal solutions without gradient instability.
- **Task Losses:** Components: Depth, shading, normal, and reflectance estimation using Berhu loss; semantic segmentation using cross-entropy and Dice loss.

Results and Conclusion

- **Datasets**: Structured3D (synthetic) and Stanford2D3D (real-world).
- **Performance on Structured3D:**

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



SCAN ME

Figure 4. Comparison with PanoFormer [1] on Structured3D, showing inferences for depth, albedo, normal, semantic, and shading. See supplementary material for larger versions and zoomed insets.

- **Performance on Stanford2D3D**
  - ***Depth Estimation:*** *Achieved impressive metrics with an MRE of 0.038, MAE of 0.065, and σ1 of 0.945.*
  - ***Semantic Segmentation****: mIoU of 54.6%, surpassing SOTA in [2]*

Conclusion

- ***MultiPanoWise*** achieves state-of-the-art performance in multi-task dense prediction for indoor panoramic images. *Future Work*: Extend to more complex tasks like inverse rendering and virtual staging.

References

1. Shen et al. PanoFormer: Panorama transformer for indoor 360 depth estimation. In Computer Vision – ECCV 2022, pages 195–211, Cham, 2022. Springer Nature.
2. Zhifeng et. Al. 360bev: Panoramic semantic mapping for indoor bird's-eye view. In 2024 IEEE/CVF Winter Conference on Applications of Computer Vision(WACV), 2024