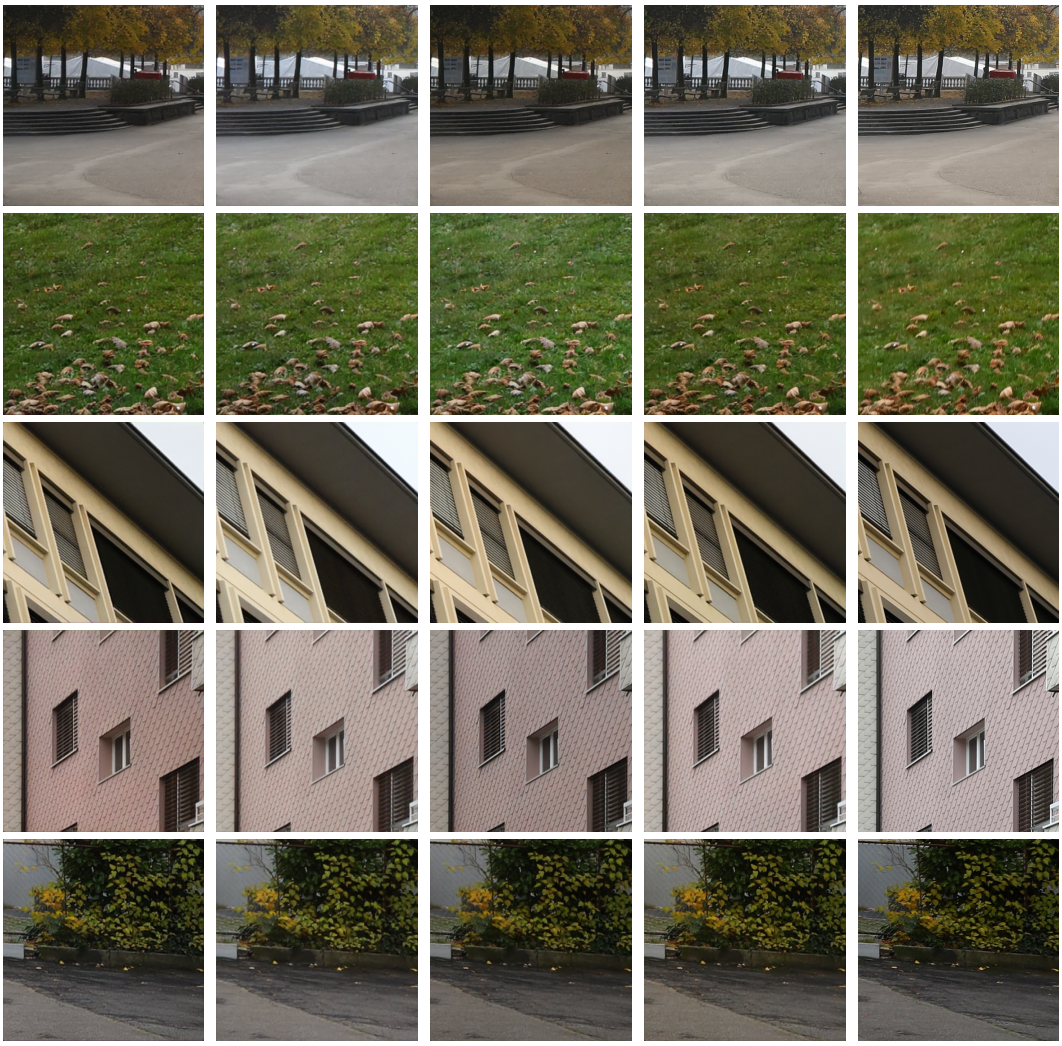
Supplementary Material for: UniISP: A Unified ISP Framework for Both Human and Machine Vision

1 RAW-TO-RGB MAPPING

A qualitative comparison of the proposed method and existing approaches on the ZRR test set is provided in fig. 1. As depicted, earlier methods including MWISPNet and AWNet exhibit difficulties in preserving color accuracy and fine-grained details. While LiteISPNet demonstrates some improvement, it remains constrained in rendering rich textures and ensuring consistent color reproduction, especially in areas with complicated illumination or fine structures. These shortcomings frequently result in noticeable artifacts or the omission of critical visual content.

By comparison, our approach yields reconstructions that closely align with the ground truth in terms of visual quality. The method faithfully reconstructs delicate patterns and subtle gradients, while achieving vibrant and color-accurate results. Exhibiting visibly sharper edges, more refined textures, and more natural color rendition, our outputs outperform those of other methods. Notably, UniISP substantially mitigates typical artifacts—such as excessive smoothing, color inaccuracies, and loss of detail—that commonly affect existing techniques.



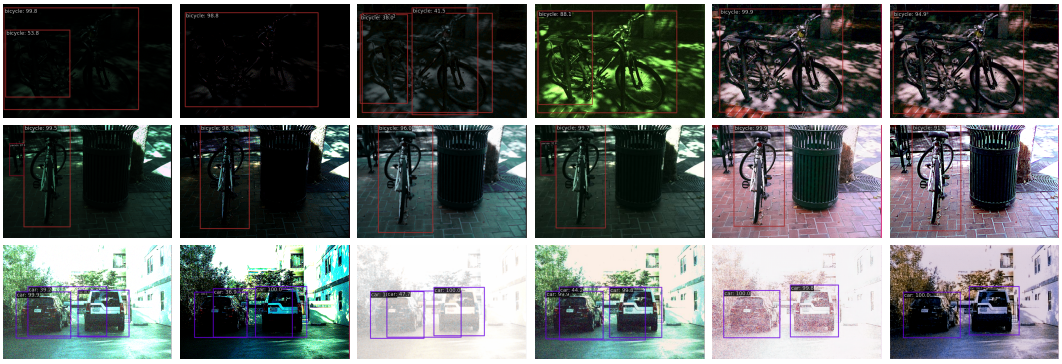
(a) MWISPNet (b) AWNet (c) LiteISPNet (d) UniISP (w/o F) (e) GT

Figure 1: Visual comparison of RAW-to-RGB results on the ZRR dataset.

2 Object Detection

As fig. 2 demonstrates, the proposed method surpasses existing approaches in both detection performance and subjective visual quality, while confirming the feasibility of simultaneously optimizing metrics for both human visual perception and machine vision.

The proposed UniISP framework uniquely optimizes RAW data processing under extremely low-light conditions for both human aesthetics and machine perception. Supervised loss guides the HAM module to ensure high-quality RAW-to-RGB outputs that align with human visual perception, while simultaneously ensuring alignment between the raw domain images and the RGB domain preferred by pre-trained backbones. Meanwhile, the Feature Adapter module leverages raw domain features to facilitate downstream detection tasks. UniISP coordinates both the HAM and FA modules to serve these dual objectives, achieving superior visual quality and robust object detection performance. The effectiveness of the method is further validated on the challenging real-world LOD dataset, as shown in fig. 3. Its generalization capability is also demonstrated across both the PASCAL RAW and LOD datasets.



(a) Demosaicing (b) InvISP (c) Karaimer *et al.* (d) RAW-Adapter (e) UniISP(w/o J) (f) UniISP

Figure 2: Visualization of object detection results on PASCAL RAW. Three rows represent dark, normal and over-exposed scenarios respectively (from top to bottom).

Default ISP InvISP RAW-Adapter UniISP

Figure 3: Experimental and visual comparisons under the real extremely dark dataset LOD.

3 Semantic Segmentation

As shown in Figure 4, our proposed UniISP consistently delivers superior semantic segmentation results across challenging illumination conditions, including dark, normal, and over-exposed RAW inputs. Compared to existing methods such as RAW-Adapter and InvISP, UniISP demonstrates remarkable robustness and generalization, accurately delineating object boundaries and achieving close semantic alignment with the ground truth. Notably, UniISP preserves fine structural details and correctly identifies small and complex objects, even under severe lighting degradation. For instance, in low-light and over-exposure scenarios, UniISP maintains clear separation of categories such as ”wall”, ”bed”, ”table”, and ”chair”, and avoids ambiguous boundaries or missing objects commonly observed in other approaches. Overall, these qualitative results underscore the effectiveness of UniISP in semantic understanding from RAW images, setting a new benchmark for scene parsing under adverse visual conditions.

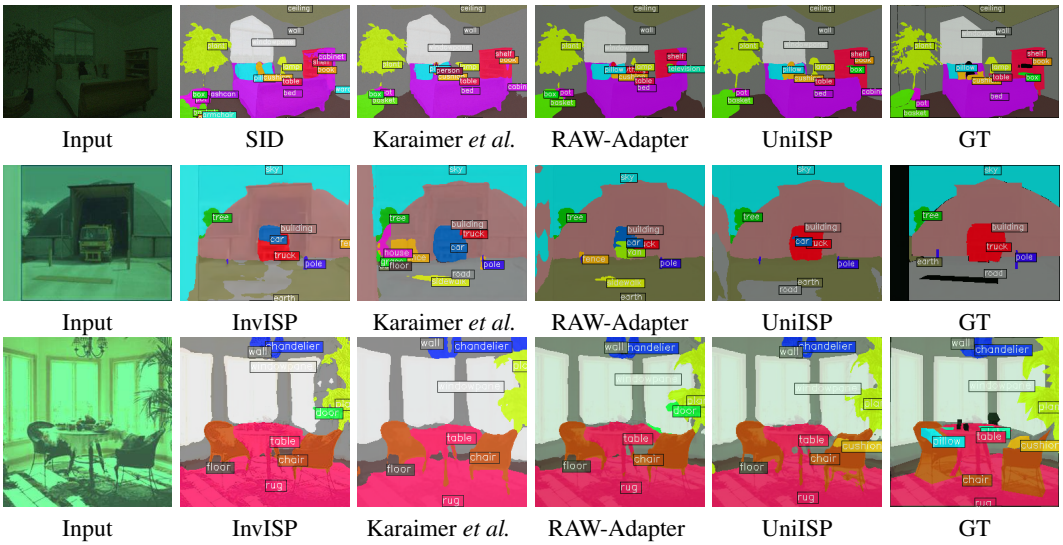


Figure 4: isualization of semantic segmentation results on ADE20K RAW. Three rows represent dark, normal and over-exposed scenarios respectively (from top to bottom).

4 LIMITATIONS

Despite the encouraging results, our work still has several limitations. First, the proposed multi-task framework requires paired RAW and RGB data for training. Although RAW data can be synthesized from RGB images, there are significant differences between synthesized RAW and genuine RAW data, both in terms of distribution and information content. Consequently, it remains challenging to acquire large-scale paired data, particularly for diverse real-world scenarios. This dependency on paired data may limit the scalability and generalization of UniISP in domains where such data is scarce or unavailable. Second, simultaneous optimization for both human perception and machine vision tasks increases training complexity, such as balancing multi-task objectives and resolving potential conflicts between perceptual and semantic requirements. Addressing these challenges may necessitate more sophisticated training strategies or adaptive loss functions. Finally, while UniISP demonstrates strong performance on the evaluated datasets, its effectiveness on other downstream applications and sensor types requires further exploration. Future work could investigate semi-supervised or unsupervised approaches to reduce the reliance on paired datasets.

5 ABLATION STUDY

The ablation study on the perceptual performance of the feature adapter for downstream tasks has been presented in table 3 and table 4. This section further validates the impact of the proposed modules on human visual perception on the ZRR dataset. As shown in table 1 (a), a systematic ablation study was conducted on the proposed HAM module and the optical flow consistency mask. PSNR and SSIM were used as evaluation metrics to analyze the contribution of each component to image quality improvement. Here, ”Base” refers to the U-Net-based baseline architecture, ”MHA” denotes the standard multi-head attention mechanism, and ”MCA” indicates the application of standard MHA in the channel dimension. The performance of MHA and MCA in the table represents the results when replacing the HAM module with them, respectively. table 1 (b) provides a detailed analysis of the impact of each component within the proposed HAM module, where RPE represents relative positional encoding, CA denotes channel attention, and SA refers to spatial attention.

Experimental results demonstrate that the proposed HAM module significantly outperforms traditional self-attention structures (MHA and MCA) in RAW-TO-RGB mapping tasks, with its effectiveness primarily stemming from the synergistic effect of channel attention (CA) and spatial attention (SA) mechanisms. Furthermore, the optical flow consistency mask mitigates alignment inaccuracies from

occlusions and homogeneous regions by selectively applying loss to reliable areas, enhancing detail preservation and color fidelity in the output.

Table 1: Ablation study results.

(a) Performance comparison of different module combinations.

Base	MHA	MCA	HAM	FlowMask	PSNR↑	SSIM↑
✓					21.37	0.8312
✓	✓				22.26	0.8421
✓		✓			22.87	0.8486
✓			✓		23.91	0.8584
✓			✓	✓	**24.14**	**0.8614**

(b) Performance analysis of the HAM module components.

	HAM			Performance	
	RPE	CA	SA	PSNR↑	SSIM↑
MCA				22.87	0.8486
MCA + RPE	✓			23.02	0.8493
MCA + CA		✓		23.64	0.8556
MCA + SA			✓	23.23	0.8513
MCA + RPE + CA	✓	✓		23.78	0.8562
MCA + RPE + SA	✓		✓	23.35	0.8521
MCA + CA + SA		✓	✓	23.85	0.8579
HAM (Ours)	✓	✓	✓	**23.91**	**0.8584**