Prefix-Guided Adaptation of Pretrained Segmenters for RGB-D Indoor Panoptic Segmentation

Anonymous Author(s)

Abstract

RGB-D indoor panoptic segmentation has seen limited progress, largely due to the scarcity of large-scale, densely annotated datasets and the poor generalization of existing methods, which often require dataset-specific retraining. To address these challenges, we propose a generalizable fine-tuning framework that efficiently adapts pretrained RGB segmentation models to the RGB-D domain using only a small, high-quality dataset (NYUv2). Specifically, we introduce two novel modules: Mutual Differential Convolution Attention (MDCA) for geometry-aware RGB-D feature fusion, and Prefix-Guided Adapter Ensemble (PGAE) for efficient modulation of decoder attention pathways without modifying the original architecture. Our approach achieves state-of-the-art panoptic segmentation performance on NYUv2, surpassing prior methods by 4.46% in Panoptic Quality (PQ). Moreover, it demonstrates strong zero-shot generalization to the significantly larger SUNRGB-D dataset, highlighting its effectiveness for scalable and transferable multi-modal scene understanding.

CCS Concepts

• **Computing methodologies** → **Image segmentation**.

Keywords

RGB-D Panoptic Segmentation, Adapter Tuning, Geometry-Aware Fusion, Pretrained Segmenters, Zero-Shot Generalization

1 Introduction

RGB-D indoor panoptic segmentation aims at holistic scene understanding by simultaneously predicting semantic categories and instance identities for every pixel. While remarkable advancements have been made on outdoor benchmarks like COCO and Cityscapes [6, 15, 30, 32], progress in indoor panoptic segmentation remains relatively limited. Indoor scenes typically contain dense arrangements of objects, severe occlusions, and significant scale variations, posing unique challenges that hinder accurate segmentation.

To better capture geometric priors inherent in indoor environments, prior works often leverage 3D representations such as truncated signed distance functions (TSDFs) [4, 7] or voxel grids [19, 27]. However, these methods are computationally intensive and sensitive to camera parameters, making them less practical for real-time high-resolution panoptic segmentation tasks. Processing RGB-D data directly in 2D image space provides a more feasible alternative, leveraging the insight that semantic categories in indoor settings exhibit relatively consistent depth distributions [19].

Nevertheless, most existing 2D RGB-D panoptic segmentation approaches, such as EMSANet [20] and EMSAFormer [10], adopt bottom-up clustering frameworks that heavily rely on hyperparameter-tuning and lack explicit global context modeling, limiting their robustness and generalizability.

In contrast, recent advances in RGB-D semantic segmentation have introduced end-to-end fusion frameworks such as CMX [31], TokenFusion [26], and GeminiFusion [13], which typically adopt dual-branch architectures trained from scratch or fully fine-tuned on RGB-D data. However, these methods often disregard pretrained RGB-only priors, leading to modality mismatch, suboptimal fusion strategies, and increased computational overhead. Geometry-aware models like DFormer [29] enhance depth utilization but remain constrained by their inability to leverage pretrained RGB representations. Meanwhile, parameter-efficient approaches such as DPLNet [8] effectively reduce tuning costs, yet exhibit limited modulation capacity, especially in deeper layers.

At the same time, recent RGB generalist segmentation models (e.g., SEEM [33] and Mask2Former [5]) exhibit strong generalization across diverse tasks, yet their adaptation to the RGB-D domain remains underexplored. Pretrained solely on RGB data, these models lack the geometric reasoning and multimodal fusion essential for indoor scenes. In contrast, RGB-D-specific models trained from scratch demand substantial computational resources and suffer from limited cross-dataset transferability.

Motivated by these limitations, we propose an efficient and generalizable fine-tuning framework that adapts pretrained RGB segmentation models to RGB-D indoor panoptic segmentation using only a small, high-quality dataset (NYUv2). The framework introduces two novel modules designed specifically to address the key challenges in RGB-D fusion and efficient adaptation:

(1) Mutual Differential Convolution Attention (MDCA): a geometry-aware RGB–D fusion mechanism that integrates RGB and depth features through bi-directional modulation and differential convolution attention, highlighting complementary cues and geometric discontinuities to enhance fine-grained spatial modeling.

(2) Prefix-Guided Adapter Ensemble (PGAE): a lightweight module that inserts prefix adapters after attention layers to inject layer-aware cues and transfer pretrained RGB attention to RGB-D tasks without altering the architecture.

Our main contributions are summarized as follows:

- We present a generalizable adaptation framework enabling efficient transfer of pretrained RGB segmentation models to RGB-D panoptic segmentation with minimal architectural modifications and high parameter efficiency.
- We introduce the Mutual Differential Convolution Attention (MDCA) module, facilitating robust geometry-aware fusion of RGB-D data via bi-directional feature modulation.
- We propose the Prefix-Guided Adapter Ensemble (PGAE), a scalable attention adaptation strategy that efficiently integrates depth cues into pretrained transformer decoders.
- Our approach achieves state-of-the-art performance on NYUv2, significantly improves upon prior methods, and demonstrates strong zero-shot transfer capability to the larger SUNRGB-D dataset.

2 Related Works

2.1 RGB-D Panoptic Segmentation

Panoptic segmentation unifies semantic and instance segmentation by assigning each pixel both a semantic label and an instance ID [14]. In the RGB-D context, most methods follow bottom-up paradigms, combining multiple prediction heads and heuristic post-processing. For instance, EMSANet [20] uses a shared encoder to jointly predict semantic labels, object centers, and embedding vectors for instance grouping. EMSAFormer [10] further incorporates Swin Transformer blocks to improve feature extraction, though it still relies on handcrafted designs and multi-branch decoders.

Sodano et al. [23] proposes a dynamic reweighting strategy to highlight informative regions, yet the final panoptic output is obtained by merging predictions from separate branches. This fragmented structure makes it difficult to enforce global consistency and hinders end-to-end optimization.

Bottom-up methods, though efficient, are sensitive to hyper-parameters and often depend on non-differentiable clustering or voting, while their modular designs limit holistic context modeling. We address these issues with an end-to-end framework that adapts a pretrained RGB panoptic segmentation model to RGB-D by injecting geometric priors and modular adapters into attention pathways, preserving generalization and ensuring efficient, robust performance in cluttered indoor scenes.

2.2 RGB-D Semantic Segmentation

RGB-D semantic segmentation enhances pixel-level predictions by combining RGB appearance with depth geometry. Many methods employ dual-branch encoders with advanced fusion strategies, such as CMX [31] and TokenFusion [26] for multi-layer token mixing, and GeminiFusion [13] for efficient pixel-wise fusion in vision transformers. These designs enable complementary modality interactions and substantially improve segmentation accuracy.

Geometry-aware approaches further leverage depth, with DFormer [29] introducing modules to guide feature learning and Liu et al. [17] proposing depth-perceptual attention to weight features by local depth reliability, proving effective in camouflaged object detection.

Nonetheless, most RGB-D models depend on full fine-tuning and heavy dual-branch networks, incurring high computational cost and limited adaptability. Lightweight schemes like DPLNet [8] reduce parameters via dual-prompt tuning but remain shallow, failing to influence deeper layers or attention dynamics and underutilizing pretrained RGB priors.

In contrast, we adapt generalist RGB segmentation models—such as SEEM [33]—to RGB-D via a prefix-guided adapter ensemble and geometry-aware fusion module, enabling parameter-efficient tuning while explicitly leveraging depth for improved feature integration.

2.3 Adapters in Vision Tasks

Adapters were first proposed in NLP for parameter-efficient transfer learning [12], allowing task adaptation via small bottleneck modules inserted into transformer layers. This idea was later extended to vision transformers (ViTs) [9] and gained traction in large-scale segmentation models. For instance, Abou Baker and Handmann [1] showed that tuning only SAM's decoder suffices for strong transfer performance, while Chen et al. [3] integrated adapter modules into both encoder and decoder for improved robustness.

Adapters have also been widely used in multimodal models. For example, Gao et al. [11] enhanced CLIP via dual-branch adapters, and Abou Baker et al. [2] applied adapter and LoRA tuning to segmentation frameworks like SEEM and MaskDINO, matching full fine-tuning with minimal parameter updates.

Recent work emphasizes not only where to insert adapters, but also how to design them for greater expressivity. Inspired by prefix-tuning [16], our approach assigns each decoder layer a distinct learnable prefix shared across its parallel adapters, yielding a lightweight design that inserts prefix-conditioned modules after encoder–decoder attention layers to modulate pretrained attention without altering the architecture.

3 Method

This section presents our modular framework for RGB-D panoptic segmentation, as shown in Figure 1. We first introduce the **Mutual Differential Convolution Attention (MDCA)** module, which novelly fuses RGB and depth features via mutual gated interaction and differential attention. Next, we describe the **Prefix-Guided Adapter Ensemble (PGAE)**, a lightweight module that adapts transformer attention to RGB-D tasks through prefix-conditioned adapters. Finally, we show how **modern decoder architectures** naturally support the integration of both modules.

3.1 Mutual Differential Convolution Attention

Let $F^{rgb} \in \mathbb{R}^{B\times C_{rgb}\times H\times W}$ and $F^{d} \in \mathbb{R}^{B\times C_{d}\times H\times W}$ denote RGB and depth features at scale k, respectively. As illustrated in Figure 2, the MDCA fusion process proceeds as follows:

1) Mutual Gated Fusion. To exploit semantic complementarity between RGB and depth features, we introduce a **bi-directional gating mechanism** for reciprocal guidance. The input features are first projected into a shared latent space via 1×1 convolutions:

$$f^{rgb} = W^{rgb}F^{rgb}, \quad f^{d} = W^{d}F^{d}. \tag{1}$$

Following the squeeze-and-excitation (SE) paradigm,

$$\mathrm{SE}(f) = \sigma\left(U_2 \cdot \phi\left(U_1 \cdot \mathrm{GAP}(f)\right)\right), \tag{2}$$

where $\mathrm{GAP}(\cdot)$ denotes global average pooling, $U_1(\cdot)$ and $U_2(\cdot)$ are projection layers for channel reduction and expansion with $U_1 \in \mathbb{R}^{\frac{C}{s}\times C}$ and $U_2 \in \mathbb{R}^{C\times\frac{C}{s}}$, $\phi(\cdot)$ is ReLU, and $\sigma(\cdot)$ is sigmoid. Here s is the reduction ratio (e.g., s=16).

The channel-wise attention from one modality is used to modulate the other:

$$g^{\mathrm{rgb}\leftarrow\mathrm{d}} = \mathrm{SE}(f^{d}), \quad g^{\mathrm{d}\leftarrow\mathrm{rgb}} = \mathrm{SE}(f^{rgb}). \tag{3}$$

The fused representation is obtained via bidirectional modulation:

$$F^{\mathrm{fuse}} = g^{\mathrm{rgb}\leftarrow\mathrm{d}} \cdot F^{rgb} + g^{\mathrm{d}\leftarrow\mathrm{rgb}} \cdot F^{d}. \tag{4}$$

This design encourages cross-guided information flow, enabling the fusion to adapt to semantic alignment across modalities.

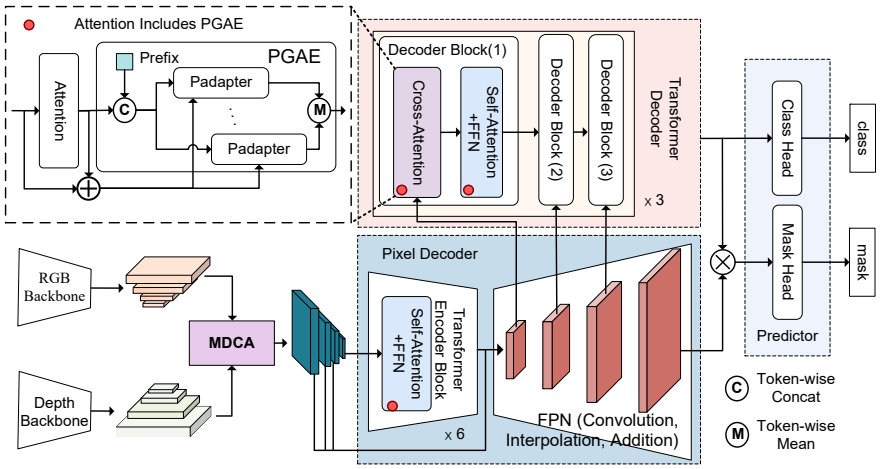

Figure 1: Overview of the proposed RGB-D panoptic segmentation framework, which adapts a generalist segmentation model to the RGB-D domain via two novel modules: MDCA and PGAE. The base architecture consists of a backbone, a pixel decoder, a transformer decoder, and a predictor. For RGB-D adaptation, the backbone is retained, MDCA fuses multi-scale RGB and depth features, and PGAE inserts prefix-conditioned adapters into the encoder and decoder attention layers to enable layer-specific attention patterns.

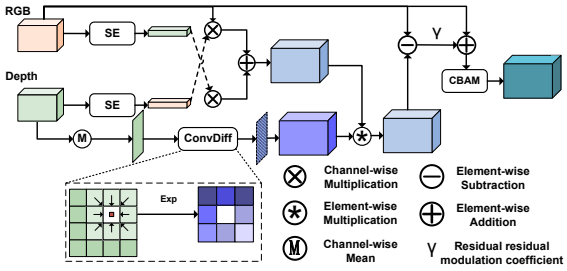

Figure 2: Architecture of the MDCA module for RGB-D feature fusion at a given scale.

2) Differential Convolution Attention. While self-attention captures global dependencies, depth cues are inherently local, where distant interactions may introduce noise rather than meaningful context. To better leverage geometric structures in depth data, we propose a **differential convolution attention mechanism** that explicitly captures local depth variations.

Let $D \in \mathbb{R}^{B\times 1\times H\times W}$ denote the input depth map (averaged across channels). For each spatial location (i, j), we compute a patch-level similarity response by comparing neighboring depths to the center pixel, weighted by an exponential kernel:

$$\delta_{i,j} = \frac{1}{K^2} \sum_{(u,v)\in\mathcal{N}_{i,j}} \exp\left(-|D(i+u, j+v) - D(i, j)|\right). \quad (5)$$

This produces a soft mask $\delta \in \mathbb{R}^{B\times 1\times H\times W}$ that reflects local structural smoothness, which is expressed as ConvDiff in Figure 2.

To extract structural cues, δ is passed through depthwise and pointwise convolutions with a sigmoid activation, yielding a multi-channel attention mask used to modulate the fused features:

$$m_{\text{geo}} = \sigma\left(\text{Conv}_{1\times 1}\left(\text{Conv}_{K\times K}.(\delta)\right)\right) \quad (6)$$

$$F^{\text{geo}} = m_{\text{geo}} \cdot F^{\text{fuse}}. \quad (7)$$

3) Residual Modulation and Attention Refinement. To adaptively balance semantic and geometric information, we blend the original RGB feature with the enhanced result:

$$F^{\text{ref}} = F^{rgb} + \gamma_k \cdot (F^{\text{geo}} - F^{rgb}). \quad (8)$$

where γ_k is a learnable scalar. The refined feature is passed through a Convolutional Block Attention Module (CBAM) to further enhance relevant content:

$$F^{MDCA} = CBAM(F^{\text{ref}}). \quad (9)$$

4) Multi-Scale Fusion Strategy. The MDCA module is applied at multiple feature scales k, with independently learned attention weights at each level. This enables fine-grained geometric enhancement at lower layers and context-aware fusion at higher levels.

3.2 Prefix-Guided Adapter Ensemble

To enable effective transfer from generalist RGB segmentation models to RGB-D panoptic tasks, we introduce the **Prefix-Guided Adapter Ensemble (PGAE)**—a lightweight, plug-and-play module designed to modulate pretrained attention patterns through learnable prompts.

PGAE builds on the hypothesis that attention maps encode rich semantic and spatial context, and their targeted adaptation is key to cross-modality generalization. Since decoder layers attend to features at different resolutions and semantic levels, we assign each layer a distinct learnable prefix to encode layer-specific inductive biases and capture resolution-aware cues, complementing the coarse-to-fine refinement of modern decoders.

PGAE integrates three key components: (1) strategic insertion into transformer decoder layers, (2) prefix-guided modulation tailored to the hierarchical roles of different layers, and (3) a parallel ensemble of adapter branches to improve representational capacity.

3.2.1 Adapter Insertion Strategy. PGAE modules are inserted after both the self-attention and cross-attention layers in each decoder block, as well as after the self-attention layers in the encoder (highlighted in red in Figure 3). For an attention layer with input $A_{in} \in \mathbb{R}^{B\times L\times D}$ and output $A_{out} \in \mathbb{R}^{B\times L\times D}$, the adapter receives two inputs: the attention output x and the residual input r.

$$x = A_{out}, \quad r = A_{in} + A_{out}. \tag{10}$$

By positioning adapters immediately after attention layers, PGAE enables the modulation of high-level relational cues while preserving the original attention structure. The residual r acts as a contextual anchor for adapter computation.

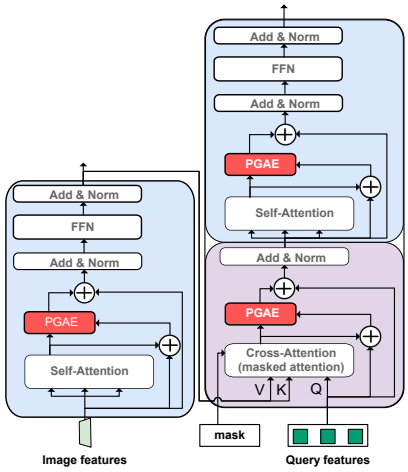

Figure 3: Illustration of PGAE insertion locations with corresponding inputs and output.

3.2.2 Prefix-Guided Residual Adapter. To accommodate the heterogeneous nature of attention maps across multi-scale decoder layers, we propose a **prefix-guided residual adapter (Padapter)** that enables layer-specific attention adaptation with minimal computational overhead.

Given the attention output $x \in \mathbb{R}^{B\times L\times D}$ and its residual input $r \in \mathbb{R}^{B\times L\times D}$, we introduce a **learnable prefix token** $P_{\text{prefix}} \in \mathbb{R}^{1\times 1\times D}$, which is broadcast along the batch dimension to obtain $P \in \mathbb{R}^{B\times 1\times D}$.

$$\tilde{x} = \text{Concat}(P, x), \quad \tilde{r} = \text{Concat}(\mathbf{0}, r), \tag{11}$$

where $\tilde{x}, \tilde{r} \in \mathbb{R}^{B\times (L+1)\times D}$, and $\mathbf{0} \in \mathbb{R}^{B\times 1\times D}$ is a zero tensor used to pad the residual input for alignment.

The concatenated inputs are then processed by N parallel prefix adapters, each following the **Pfeiffer-style residual adapter** formulation [18]:

$$o_i = \text{Adapter}_i(\tilde{x}, \tilde{r}) = \tilde{r} + W_u^{(i)} \,\text{ReLU}(W_d^{(i)} \tilde{x}), \quad i = 1, \ldots, N. \tag{12}$$

where $W_d^{(i)} \in \mathbb{R}^{d_a\times D}$ and $W_u^{(i)} \in \mathbb{R}^{D\times d_a}$ are learnable projection matrices, and d_a denotes the adapter bottleneck dimension.

Each adapter output o_i is then split into two parts: prefix component o_i^P and token component o_i^x. For **prefix-guided modulation**, o_i^P is broadcast and concatenated with o_i^x along the channel dimension to form a fused representation, which is processed by a shared two-layer MLP with residual connection:

$$y_i = o_i^x + W_2 \cdot \text{GELU}\left(W_1 \cdot \left[o_i^p;\, o_i^x\right]\right), \tag{13}$$

where $W_1 \in \mathbb{R}^{2D\times d_m}$ and $W_2 \in \mathbb{R}^{d_m\times D}$ are shared across all adapters, and d_m denotes the hidden dimension of the MLP. The residual path ensures stable feature modulation while preserving the semantics of the original token features.

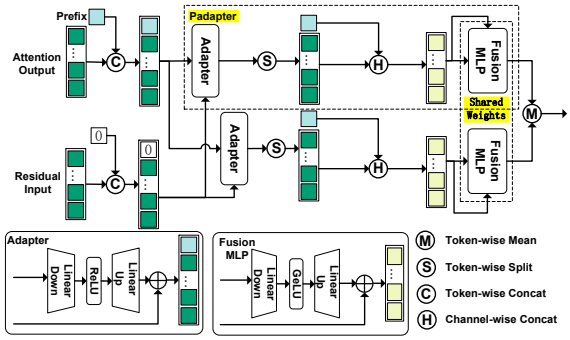

Figure 4: Architecture of the PGAE module.

3.2.3 Adapter Ensemble and Aggregation. To increase the modeling capacity of PGAE, we deploy an ensemble of N parallel Padapter, all operating on the same input tuple $(\tilde{x}, \tilde{r})$. Each Padapter branch $\mathcal{A}_i$ is independently parameterized, allowing diverse transformation behaviors to emerge under **a shared prefix context**:

$$y_i = \mathcal{A}_i(\tilde{x}, \tilde{r}), \quad i = 1, \ldots, N. \tag{14}$$

The final output is averaged to ensure smooth aggregation across diverse adapter branches, a strategy that balances performance and efficiency:

$$y = \frac{1}{N}\sum_{i=1}^{N} y_i. \tag{15}$$

3.3 Compatibility with Modern Segmentation Decoders

Our proposed modules, PGAE and MDCA, are inherently compatible with the architectural design of modern transformer-based segmentation decoders, enabling seamless integration without modifying core architectures. This compatibility arises from two key architectural characteristics:

Foreground-Guided Masked Attention. Modern Transformer decoders, such as Mask2Former [5], typically employ masked cross-attention that prioritizes foreground regions identified by earlier decoder layers. In these designs, cross-attention precedes self-attention, collecting spatially relevant features first before global reasoning. PGAE naturally complements this structure by introducing learnable prefix prompts, guiding attention along semantically meaningful paths, thus enhancing foreground-focused interactions and minimizing background interference.

Layer-wise Auxiliary Supervision with Decoupled Heads. Modern decoders decouple class and mask embeddings, allowing each layer to independently predict segmentation masks and contribute to the overall loss. MDCA leverages this property by injecting multi-scale RGB-D fusion into specific decoder layers under intermediate supervision, reinforcing hierarchical consistency and improving optimization efficiency.

4 Experiments

4.1 Datasets and Evaluation Metrics

The NYUv2 dataset [22] comprises 1,449 RGB-D indoor images with dense annotations for panoptic and semantic segmentation. Following standard protocol, we use 795 images for training and 654 for testing, focusing on the 40 most frequent semantic categories. Structural elements (e.g., wall, floor, ceiling) are treated as stuff, while other objects are treated as things, totaling approximately 12,092 and 9,874 annotated instances in the training and test splits, respectively.

SUNRGB-D [24] includes 10,335 RGB-D images with semantic labels for 37 NYUv2 categories, split into 5,285 for training and 5,050 for testing. While suitable for semantic segmentation, SUNRGB-D's instance annotations—even after reconstruction from 3D boxes [20]—remain too sparse and incomplete to support reliable or meaningful panoptic segmentation evaluation.

We evaluate panoptic segmentation using three standard metrics—Panoptic Quality (PQ), Segmentation Quality (SQ), and Recognition Quality (RQ)—following Seichter et al. [20]. For semantic segmentation, we report mean Intersection over Union (mIoU).

4.2 Implementation Details

Experimental Setup. We build our framework upon SEEM [33], using Focal-L and Focal-T [28] as RGB and depth backbones, respectively. The pixel decoder adopts the Feature Pyramid Network (FPN) version, and all components are initialized from official SEEM checkpoints.

Panoptic segmentation serves as our primary training and evaluation task, conducted on the NYUv2 dataset.

The model is trained on NYUv2 for 50 epochs using a single NVIDIA RTX A6000 GPU with AdamW optimizer and a base learning rate of 1×10^{-4}. We employ a cosine warm-up schedule, mixed-precision (FP16) training, and a batch size of 16. Input images are resized with the shorter side to 512 pixels (max 1024), and augmented via random scaling, cropping, and flipping before being cropped to 512×1024.

Ablation Settings. Unless specified, all ablations follow the default setup: most parameters are frozen, updating only mask and class embeddings; adapters and fusion modules are toggled as needed. Built on SEEM with RGB-D adjustments, PGAE uses prefix length L_p=1, two parallel adapters (N=2) with reduction ratio 4, and a Fusion MLP with a reduction ratio of 32; prefixes are trainable. MDCA applies K=3 differential convolution for balanced geometry–context modeling. **All visual results use this configuration.**

4.3 Comparison to State-of-the-art Methods

We evaluate our method on the NYUv2 dataset and achieve **state-of-the-art** performance across all metrics (Table 1). Our model reaches a PQ of **55.61**, outperforming the previous best method by **4.46** points, with consistent improvements in RQ, SQ, and a mIoU of **62.38**. These results demonstrate the strength of MDCA and PGAE in enhancing instance discrimination and global scene understanding in complex indoor environments through cross-modal fusion and layer-aware adaptation.

Table 1: Panoptic segmentation performance on NYUv2.

Method	PQ (%) ↑	RQ (%) ↑	SQ (%) ↑	mIoU(%) ↑
EMSANet [20]	47.38	55.95	83.74	53.80
EMSANet (PanopticNDT) [21]	51.15	59.59	84.80	59.02
Ours	**55.61**	**63.74**	**86.62**	**62.38**

4.4 Generalizablity Evaluation

To ensure fair comparison with a broader range of methods, we derive semantic masks from our model's panoptic outputs on NYUv2, leveraging the fact that panoptic segmentation assigns a semantic label to every pixel. As shown in Table 2, our method achieves the highest mIoU of **62.38**.

Table 2: Semantic segmentation on NYUv2.

Method	mIoU(%) ↑
DFormer [29]	57.20
EMSANet (PanopticNDT) [21]	59.02
DPLNet [8]	59.30
OmniVec [25]	60.80
GeminiFusion [13]	60.90
Ours	**62.38**

To further assess generalization, we conduct a cross-dataset evaluation: the model is trained exclusively on NYUV2 and directly tested on the significantly larger SUNRGB-D dataset (approximately 7× larger). As SUNRGB-D lacks full panoptic annotations, we evaluate performance using its semantic labels. As shown in Table 3, our method also achieves the best results.

Table 3: Zero-shot semantic segmentation on SUNRGB-D.

Method	mIoU(%) ↑
DFormer [29]	52.50
DPLNet [8]	52.80
GeminiFusion [13]	53.30
Ours	**53.47**

4.5 Ablation Study

To assess the effectiveness of MDCA and PGAE, we perform ablation studies by replacing each with a strong alternative. Specifically, PGAE is substituted with **SAdapterE** [2], which employs stacked residual adapters with shared inputs; both are inserted at the same locations with two adapters for fair comparison. MDCA is replaced with the **Depth-weighted Cross-attention Fusion (DCF)** module [17], which adaptively weights RGB and depth features based on the estimated depth reliability. Originally developed for camouflaged object detection, DCF has demonstrated strong fusion performance.

As shown in Table 4, removing either MDCA or PGAE leads to performance drops, confirming their complementary roles. Replacing MDCA with DCF reduces PQ to 54.35, while substituting PGAE with SAdapterE lowers mIoU to 61.84.

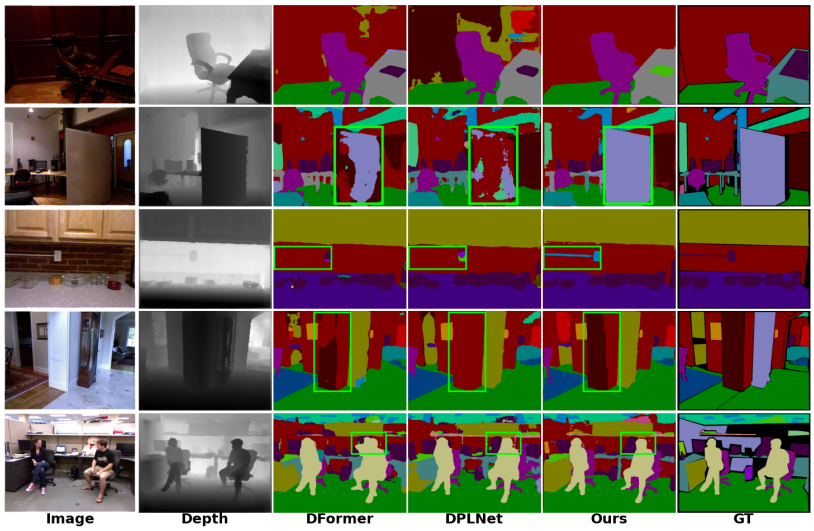

Figure 5: Qualitative comparison of semantic segmentation results.

Table 4: Ablation study of MDCA and PGAE on NYUv2.

Method	Trained Params	**PQ (%) ↑**	**mIoU (%) ↑**
w/o MDCA + PGAE	14.76M (3.84%)	55.215	60.15
w/o MDCA	16.04M (4.16%)	54.347	61.67
w/o PGAE	16.39M (4.25%)	55.258	61.84
Ours (MDCA + PGAE)	17.67M (4.57%)	**55.610**	**62.38**

To validate the efficiency of **PGAE**, we evaluate a range of fine-tuning strategies, as summarized in Table 5. PGAE(2) achieves 55.61 PQ and a state-of-the-art mIoU of 62.38 using only 4.27% of parameters—outperforming both SAdapter variants and full decoder tuning in semantic accuracy, and demonstrating strong generalization. Although full head tuning yields the highest PQ, it requires 134M (M denotes million) parameters, highlighting the favorable trade-off offered by PGAE. The proportion of trainable parameters is defined as $\frac{100\times N_{\text{train}}}{N_{\text{base}}+N_{\text{adapt}}}$.

Table 5: Comparison of fine-tuning efficiency using PGAE on NYUv2.

Tuning Module	Trained Params	**PQ (%) ↑**	**mIoU (%) ↑**
Full Head	134.61M (35.49%)	**58.438**	61.18
Only Decoder	111.44M (29.38%)	56.618	60.60
Only CE & ME	10.07M (2.65%)	49.787	60.19
SAdapterE (2)	16.39M (4.25%)	55.017	61.76
SAdapterE (4)	22.71M (5.79%)	55.819	62.12
SAdapterE (8)	33.73M (8.37%)	56.334	61.84
PGAE (2, default)	17.67M (4.57%)	55.613	**62.38**
PGAE (4)	23.99M (6.81%)	55.226	61.83
PGAE (8)	36.63M (9.03%)	56.651	62.02

We further investigate the effect of adapter scaling. PGAE(2), which employs mean aggregation, offers the best balance of efficiency and performance. PGAE(4) underperforms due to redundancy, whereas PGAE(8) regains effectiveness by increasing capacity and attention specialization—highlighting its potential for richer multimodal modeling.

As shown in Table 6, our method achieves superior performance within just 50 epochs, outperforming RGB-D segmentation baselines like DFormer and DPLNet trained for 500 epochs. This rapid convergence stems from two key factors: (1) pretrained RGB segmentation provides strong initialization for RGB-D adaptation; and (2) masked attention focuses learning on foreground regions, suppressing background noise and accelerating optimization. Additionally, PGAE enables efficient attention modulation with minimal parameter overhead, enhancing both accuracy and training efficiency.

Table 6: Comparison of training efficiency on NYUv2.

Method	Epoch	**Best mIoU (%) ↑**
DFormer	50	51.38
DPLNet	50	56.60
DFormer	500	57.20
DPLNet	500	59.30
Ours	50	**62.38**

4.6 Visual Comparison

As shown in Figure 5, our method consistently outperforms DFormer and DPLNet, yielding coherent masks under occlusion, smooth large-object segmentation, and fine details for small or transparent structures. It better preserves planar geometry through implicit 2.5D scene understanding and generates instance masks that closely follow object contours, particularly for human silhouettes. These results demonstrate the strong transferability and geometric awareness enabled by MDCA and PGAE.

5 Conclusion

We propose an adaptation framework that extends the generalist segmentation model SEEM to RGB-D indoor panoptic segmentation for the first time. By effectively leveraging the pretrained model's rich visual priors, our approach achieves state-of-the-art performance on the NYUv2 dataset with minimal training overhead. In addition to outperforming task-specific baselines in both accuracy and training efficiency, it demonstrates strong zero-shot generalization to the larger SUNRGB-D dataset. The framework preserves the scalability and modularity of the original architecture, offering a practical and versatile solution for multi-modal scene understanding.

References

[1] Nermeen Abou Baker and Uwe Handmann. 2023. Don't waste SAM. In *The 31th European Symposium on Artificial Neural Networks (ESANN 2023)*. 429–434.

[2] Nermeen Abou Baker, David Rohrschneider, and Uwe Handmann. 2024. Parameter-Efficient Fine-Tuning of Large Pretrained Models for Instance Segmentation Tasks. *Machine Learning and Knowledge Extraction* 6, 4 (2024), 2783–2807.

[3] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. 2023. Sam-adapter: Adapting segment anything in underperformed scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3367–3375.

[4] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. 2023. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21674–21683.

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1290–1299.

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. 2021. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems* 34 (2021), 17864–17875.

[7] Jaesung Choe, Sunghoon Im, Francois Rameau, Minjun Kang, and In So Kweon. 2021. Volumefusion: Deep depth fusion for 3d scene reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*. 16086–16095.

[8] Shaohua Dong, Yunhe Feng, Qing Yang, Yan Huang, Dongfang Liu, and Heng Fan. 2024. Efficient multimodal semantic segmentation via dual-prompt learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 14196–14203.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. [n. d.]. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

[10] Söhnke Benedikt Fischedick, Daniel Seichter, Robin Schmidt, Leonard Rabes, and Horst-Michael Gross. 2023. Efficient multi-task scene analysis with rgb-d transformers. In *2023 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–10.

[11] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2024. Clip-adapter: Better vision-language models with feature adapters. *International Journal of Computer Vision* 132, 2 (2024), 581–595.

[12] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *International conference on machine learning*. PMLR, 2790–2799.

[13] Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. 2024. GeminiFusion: efficient pixel-wise multimodal fusion for vision transformer. In *Proceedings of the 41st International Conference on Machine Learning*. 21753–21767.

[14] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollár. 2019. Panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 9404–9413.

[15] Wenbo Li, Zhicheng Zhang, Enze Xie, Zhaoxin Yu, Wenhai Wang, Tong Lu, Ping Luo, and Jifeng Dai. 2022. Panoptic SegFormer: Delving Deeper into Panoptic Segmentation with Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190* (2021).

[17] Xinran Liu, Lin Qi, Yuxuan Song, and Qi Wen. 2024. Depth awakens: A depth-perceptual attention fusion network for RGB-D camouflaged object detection. *Image and Vision Computing* 143 (2024), 104924.

[18] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulic, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. 2020. AdapterHub: A Framework for Adapting Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, Qun Liu and David Schlangen (Eds.). Association for Computational Linguistics, 46–54. doi:10.18653/V1/2020.EMNLP-DEMOS.7

[19] Trung T Pham, Thanh-Toan Do, Niko Sünderhauf, and Ian Reid. 2018. Scenecut: Joint geometric and object segmentation for indoor scenes. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 3213–3220.

[20] Daniel Seichter, Söhnke Benedikt Fischedick, Mona Köhler, and Horst-Michael Groß. 2022. Efficient multi-task rgb-d scene analysis for indoor environments. In *2022 International joint conference on neural networks (IJCNN)*. IEEE, 1–10.

[21] Daniel Seichter, Benedict Stephan, Söhnke Benedikt Fischedick, Steffen Mueller, Leonard Rabes, and Horst-Michael Gross. 2023. PanopticNDT: Efficient and robust panoptic mapping. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7233–7240.

[22] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*. Springer, 746–760.

[23] Matteo Sodano, Federico Magistri, Tiziano Guadagnino, Jens Behley, and Cyrill Stachniss. 2023. Robust double-encoder network for rgb-d panoptic segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4953–4959.

[24] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 567–576.

[25] Siddharth Srivastava and Gaurav Sharma. 2024. Omnivec: Learning robust representations with cross modal sharing. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1236–1248.

[26] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. 2022. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12186–12195.

[27] Haihong Xiao, Hongbin Xu, Wenxiong Kang, and Yuqiong Li. 2024. Instance-aware monocular 3D semantic scene completion. *IEEE Transactions on Intelligent Transportation Systems* 25, 7 (2024), 6543–6554.

[28] Jianwei Yang, Chunyuan Li, Xiyang Dai, and Jianfeng Gao. 2022. Focal modulation networks. *Advances in Neural Information Processing Systems* 35 (2022), 4203–4217.

[29] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. 2024. DFormer: Rethinking RGBD Representation Learning for Semantic Segmentation. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=h1sFUGlI09

[30] Qihang Yu, Huiyu Wang, Siyuan Qiao, Maxwell Collins, Yukun Zhu, Hartwig Adam, Alan Yuille, and Liang-Chieh Chen. 2022. k-means Mask Transformer. In *European Conference on Computer Vision*. Springer, 288–307.

[31] Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. 2023. CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems* 24, 12 (2023), 14679–14694.

[32] Xueyan Zou, Zi-Yi Dou, Jianwei Yang, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Jianfeng Wang, Lu Yuan, et al. 2023. Generalized decoding for pixel, image, and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 15116–15127.

[33] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao, and Yong Jae Lee. 2023. Segment everything everywhere all at once. *Advances in neural information processing systems* 36 (2023), 19769–19782.