Efficient Indoor Robotic Scene Analysis with Encoder-Centric Vision Transformers

Söhnke Benedikt Fischedick, Benedict Stephan, Daniel Seichter, and Horst-Michael Gross
Ilmenau University of Technology, Neuroinformatics and Cognitive Robotics Lab
98684 Ilmenau, Germany
soehnke.fischedick@tu-ilmenau.de, ORCID: 0000-0001-8447-0584

***Abstract*—Indoor mobile robots require a comprehensive understanding of their surroundings to navigate and interact safely in cluttered environments. This typically involves multiple complementary perception tasks, such as scene classification, semantic segmentation, and panoptic segmentation, under real-time constraints. While recent mask-based segmentation methods already shift instance separation into the model, many existing approaches still remain computationally expensive and have mainly been studied in RGB-only settings. In this paper, we propose IRSAFormer, a Transformer-based approach for indoor scene analysis that follows the efficient Encoder-only Mask Transformer (EoMT) paradigm by injecting query tokens into a largely unmodified Vision Transformer (ViT) backbone and using lightweight task heads directly on top of it. To address the limitations of RGB-only perception, we analyze RGB-D incorporation strategies. Our results indicate that depth provides clear gains when pre-training is limited, whereas strong large-scale pre-training narrows the gap between RGB and RGB-D. Beyond closed-set prediction, we investigate distilling text-aligned embeddings as a more flexible output representation for language-driven downstream applications. We find that scene-level embeddings distill reliably, whereas query token-level embeddings remain more challenging. Overall, IRSAFormer achieves state-of-the-art performance on NYUv2, SUN RGB-D, and ScanNet V2 while maintaining high efficiency, reaching 26.8 FPS on an NVIDIA Jetson AGX Orin with a ViT-Base backbone.**

I. Introduction

Mobile robots need a detailed understanding of their surroundings to operate autonomously in indoor environments. This understanding must capture both *what* is present in the scene and *where* it is located–from room layout down to individual objects. For example, when a mobile robot is instructed to *"drive to the chair in the living room"*, it must identify the room, distinguish between multiple object instances, and select the correct target based on its semantics.

In practice, this requires combining several perception tasks. Scene classification [1] provides room-level context, semantic segmentation yields dense, per-pixel class labels, and panoptic segmentation [2] additionally separates object instances. Many existing scene understanding approaches rely on encoder-decoder architectures that operate on dense pixel-level predictions [1], [3]–[5]. While this formulation is effective for semantic segmentation, panoptic segmentation typically requires additional grouping and post-processing steps, increasing overall design complexity. Recent mask-based segmentation methods offer an alternative formulation that represents objects and regions with a fixed set of query tokens [6]–[8]. Each token is decoded into a binary mask and a semantic or panoptic label, shifting instance separation into the learned model. While mask-based methods can be accurate, they often rely on specialized and computationally expensive additional network components. However, recent work on Encoder-only Mask Transformers (EoMT) [9] has shown that such network components can be avoided when the backbone is pre-trained with large-scale data and strong supervision. The key idea of EoMT is to inject query tokens into a largely unmodified Vision Transformer (ViT) [10] and attach only lightweight task heads for mask and label prediction.

In this paper, we adapt the EoMT paradigm to indoor robotic scene analysis and refer to our approach as Indoor Robotic Scene Analysis Transformer (IRSAFormer). IRSAFormer extends the EoMT paradigm to jointly address all aforementioned perception tasks–scene classification as well as semantic and panoptic segmentation–as shown in Fig. 1.

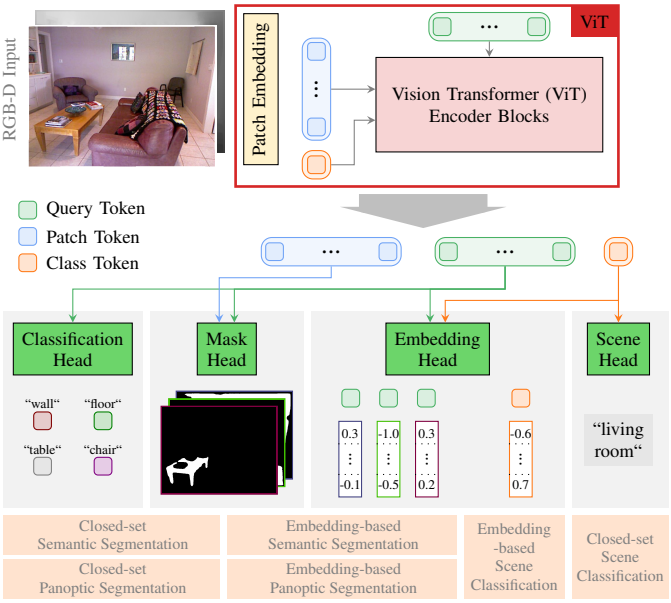

Fig. 1. Overview of our proposed Transformer-based approach for indoor robotic scene analysis (IRSAFormer). An RGB, or optional RGB-D, image is encoded into patch tokens by a Vision Transformer backbone, while additional query tokens are injected to represent objects and regions. Task-specific lightweight heads decode the resulting tokens into semantic/panoptic segmentation predictions, optional text-aligned visual embeddings, or a scene label, depending on the task configuration used during training.

This work has received funding from Carl-Zeiss-Stiftung to the project *Co-Presence of Humans and Interactive Companions for Seniors* (CO-HUMANICS).

Such indoor perception tasks are known to often benefit from additional depth cues [11], [12]. Depth provides complementary geometric information that is particularly valuable in cluttered indoor environments and challenging lighting conditions. However, the existing EoMT paradigm is limited to RGB inputs. Given the modern paradigm with a large-scale pre-trained backbone, it remains an open question how beneficial depth information still is and how it can be efficiently integrated into a token-driven, mask-based approach. We therefore compare early fusion via a 4-channel RGB-D input with dual-stream designs employing cross-modality fusion and analyze their impact across different levels of pre-training. Our results indicate that incorporating depth provides clear gains for limited pre-training configurations, whereas large-scale pre-trained backbones substantially narrow the gap between RGB and RGB-D, making RGB-only models highly competitive for indoor robotic applications. However, the competitiveness of such RGB-only models is closely tied to the availability of such checkpoints in downstream applications.

Another limitation of the EoMT paradigm is its restriction to closed-set segmentation, where predictions are limited to a predefined set of semantic classes. This constraint can hinder downstream robotic applications. For example, when specifying a target position for a mobile robot, it would be desirable to refer to objects using natural-language identifiers that go beyond a fixed set of predefined classes. Motivated by prior work [13], [14], we therefore also investigate distilling text-aligned visual embeddings as an alternative output representation to closed-set labels. Our experiments indicate that scene-level embeddings can be distilled reliably, while token-level alignment remains more challenging.

We evaluate IRSAFormer on NYUv2 [15], SUN RGB-D [16], and ScanNet V2 [17] and demonstrate state-of-the-art performance while maintaining high computational efficiency. On an NVIDIA Jetson AGX Orin, our ViT-Base configuration reaches $26.8\,\text{FPS}$, satisfying our real-time target of $\geq 25\,\text{FPS}$, while smaller configurations reach up to $55.3\,\text{FPS}$.

In summary, our main contributions are:

- an encoder-centric indoor scene analysis model (IRSAFormer) that jointly addresses scene classification, semantic segmentation, and panoptic segmentation under real-time constraints;
- a systematic analysis of RGB-D incorporation strategies for token-based transformer models, studying performance, robustness to illumination changes, and throughput trade-offs across different pre-training configurations, fusion strategies, and model scales;
- an investigation of text-aligned visual embeddings as an alternative output representation to closed-set labels.

We share the full training pipeline and model weights at GitHub: `https://github.com/TUI-NICR/IRSAFormer`.

II. Related Work

In the following, we review related work on indoor scene understanding tasks, large-scale pre-training, RGB-D processing, and text-aligned representations.

A. Indoor Scene Understanding Tasks

Indoor robotic scene understanding is commonly addressed through a combination of tasks. Scene classification assigns a single room label, such as *kitchen* or *living room*, to an input image and is commonly implemented by classifying a global descriptor obtained by pooling spatial features or by using the Vision Transformer's class token [10].

Semantic segmentation predicts a semantic label for each pixel and is often addressed using encoder-decoder architectures for dense prediction [5], [12], [18]. Beyond pixel-wise classification, recent work also formulates segmentation as mask-wise prediction, where a fixed number of queries are used to predict segments that are assigned to semantic classes [6], [7]. While mask-based methods achieve high accuracy, they often rely on specialized and computationally expensive network components, which can limit their suitability for real-time, resource-constrained applications.

Panoptic segmentation unifies semantic and instance segmentation by assigning each pixel a semantic label and, for *thing* classes (countable objects), an instance ID, whereas *stuff* classes (often background) do not require instances [2]. Top-down approaches often build on instance segmentation methods such as Mask R-CNN [19] and add semantic prediction and merging strategies [20], [21]. Bottom-up approaches extend dense prediction and retrieve instances through grouping, e.g., by predicting instance centers and per-pixel offsets and assigning pixels in a heuristic post-processing step [1], [3], [22]. Overall, bottom-up designs are common in real-time robotic applications due to their efficiency; however, panoptic prediction relies on grouping and merging steps that are not learned jointly with the network. Similar to semantic segmentation, panoptic segmentation has also increasingly been addressed using mask-based approaches that predict a fixed set of segments using query tokens [6]–[8], [23]. Each query token is decoded into a binary instance-aware mask and a corresponding semantic label, thereby enabling instance prediction without explicit pixel grouping. However, similar to semantic segmentation, most mask-based methods rely on specialized network components with high computational cost. Encoder-only Mask Transformer (EoMT) [9] reduces this complexity by processing query tokens directly within a largely unmodified Vision Transformer backbone and predicting masks and classes using lightweight task heads.

B. Large-Scale Pre-Training

Vision backbones were traditionally pre-trained in a supervised manner on ImageNet-1K [24], learning to classify images (˜1.3M images) into a fixed set of classes. This provides a strong initialization that can be fine-tuned for downstream tasks and often improves data efficiency compared to training from scratch. A common scaling step is ImageNet-21K (˜14M images), which increases label coverage but can be harder to optimize due to the limited number of examples for many classes. More recent approaches reduce the need for manual labels and enable training on much larger datasets by using objectives that do not require class annotations.

Self-supervised methods such as the DINO family [25], [26] learn from unlabeled data (˜1.7B images for DINOv3) by enforcing consistency between different augmented views of the same image and often yield features that transfer well to dense prediction. Vision-language models such as CLIP [27] learn from image-text pairs (˜400M pairs) by aligning visual and textual embeddings, providing semantically rich visual features that can be queried using text.

Knowledge distillation-based pre-training uses features of a teacher model as targets, providing a richer learning objective than one-hot class labels and often improving generalization [28]. EVA-02 [29] is a representative example that distills from a CLIP-style teacher, which provides supervision beyond the discrete ImageNet-21K label space and can yield strong downstream performance compared to purely supervised ImageNet-21K pre-training.

These modern pre-training regimes are typically computationally expensive to reproduce. As a result, many approaches adopt available pre-trained backbones and focus on adapting lightweight, task-specific components, rather than training custom encoders from scratch [9], [30].

C. RGB-D Processing

Most approaches operate on RGB, while indoor perception often benefits from additional depth cues. A common way to incorporate depth is to use a separate encoder and to fuse features at selected stages [1], [4], [5]. Such dual-stream approaches can be effective but introduce inductive biases regarding the fusion location and mechanism and add computational overhead depending on the backbone. Other methods fuse both modalities earlier, e.g., by using depth-aware convolutions or by jointly processing RGB-D in a single encoder [31]–[33]. Transformer-based work also explores token-level fusion by treating depth as an additional token stream. MultiMAE [34] uses modality-specific patch embeddings and combines RGB and depth tokens in a shared transformer, with modality-aware masking during pre-training and modality-specific decoders for downstream tasks. This avoids explicit fusion modules but increases the number of tokens processed by attention mechanism, which can become a bottleneck at higher resolutions. Depth integration has also been studied in the context of large-scale pretrained backbones. For example, recent work using DINOv2-based encoders reports that dual-stream designs with intermediate fusion can outperform simple early fusion [35].

D. Text-Aligned Embedding Representations

Beyond discrete labels for tasks such as semantic and panoptic segmentation, several works explore text-aligned embeddings as an alternative output representation for regions or entire scenes. Such embeddings can support downstream robotics applications such as retrieval [14] or mapping [36], [37]. A key challenge is obtaining localized embeddings, since many vision-language models, such as CLIP [27], provide mainly global image descriptors. Prior work follows different strategies to obtain localized text-aligned features from CLIP-style models. Some methods modify or augment CLIP features for dense or region-wise segmentation, often using additional mechanisms to produce pixel- or mask-level predictions [13], [38]. Other work uses distillation [28] or pseudo-labels to transfer supervision from CLIP-based predictors to efficient segmentation models [13]. Related distillation pipelines avoid modifying CLIP and use Alpha-CLIP [39] instead to produce region embeddings conditioned on an alpha mask. As an alternative, dino.txt [40] aligns a text encoder to a frozen DINO backbone to obtain text alignment without re-training the vision encoder but still requires large-scale paired data to train a text encoder from scratch.

III. Efficient Indoor Robotic Scene Analysis with Encoder-Centric Vision Transformers

Our Transformer-based approach for indoor robotic scene analysis (IRSAFormer) is illustrated in Fig. 2. We adopt the EoMT-style encoder-centric design [9] and extend it to jointly address multiple prediction tasks as well as RGB-D processing and embedding estimation. At a high level, a Vision Transformer (ViT) [10] backbone processes a unified token stream composed of multiple tokens, followed by lightweight heads that decode the tokens to produce scene-level, semantic, and panoptic predictions. This design targets joint prediction with minimal backbone changes and efficient inference.

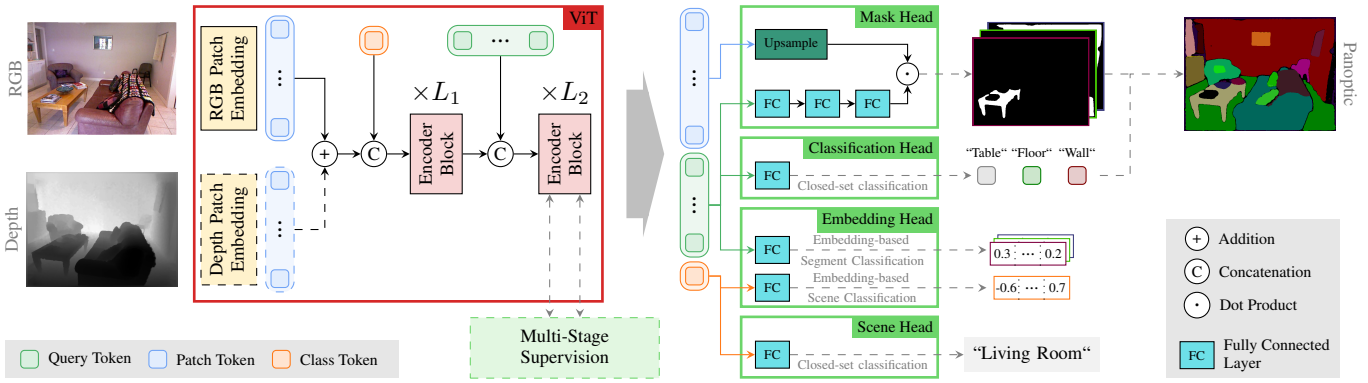

Fig. 2. Architecture of our proposed Transformer-based approach for indoor robotic scene analysis (IRSAFormer). It adopts an EoMT-style encoder-centric design [9] and extends it to jointly address multiple prediction tasks, RGB-D processing, and embedding estimation. A Vision Transformer (ViT) [10] backbone processes a unified token stream composed of patch-embedding tokens extracted from an RGB input and, optionally, a depth input, together with a class token and a fixed set of injected query tokens. Subsequent lightweight heads decode the tokens to produce scene-level, semantic, and panoptic predictions.

A. Token Construction and Processing

Given an RGB image and an optional depth image, modality-specific patch embeddings are computed and combined into a single patch-token stream. As commonly done in ViTs, a learnable class token is included alongside the patch tokens. After L_1 backbone blocks, K learnable query tokens are injected and processed together with the token stream for the remaining L_2 blocks.

Subsequent lightweight heads decode the tokens into predictions. In the mask head, each query token is decoded into a binary mask that specifies *where* a single object instance or a stuff region is located in the image. The architecture follows the EoMT [9] design. It projects each query token to a mask-specific embedding and combines it with upsampled feature maps derived from the patch tokens to produce dense mask logits. The resulting K masks represent a common set of regions and are reused across all mask-level tasks. At the same time, the query tokens are processed in the classification head to predict *what* the regions represent in a closed set of semantic classes. The scene head follows a similar classification design but uses the class token to predict *what* the overall scene represents at the room level.

In addition to closed-set labels, we consider text-aligned visual embeddings as an alternative output representation. For mask-level descriptors, we replace the classification head with an embedding head that maps each query token to a D-dimensional vector. For scene-level descriptors, an independent branch in the embedding head maps the class token to a vector of the same dimensionality. We treat embedding prediction as an optional output head and evaluate it separately from closed-set tasks. Compared to per-pixel embedding maps of DVEFormer [14], our token-based formulation predicts a single embedding per region as an output. This reduces memory, as we only store one vector per query, and the number of queries is fixed and typically far smaller than the number of pixels in a dense embedding map.

B. Depth as Additional Input Modality

Depth provides complementary geometric information that can improve indoor perception tasks [11], [12]. For IR-SAFormer, we study two depth incorporation strategies that preserve the encoder-centric structure of our approach.

In a single-stream early-fusion design (shown in Fig. 2), depth is incorporated by adding depth patch embeddings to the RGB patch embeddings, such that RGB and depth are processed in a single token stream with minimal overhead compared to RGB-only processing.

In a dual-stream design, RGB and depth are processed by separate backbones and periodically fused using a lightweight squeeze-and-excitation [41] fusion module, as commonly done in RGB-D approaches [1], [12], [35]. We apply fusion at selected stages, specifically after every three encoder blocks, and study both fusion directions by injecting features either into the RGB stream or into the depth stream. After fusion, query tokens are applied to the fused stream, and all prediction heads remain unchanged. Compared to early fusion, this design increases computational cost but can preserve modality-specific structure before fusion. We intentionally avoid joint token-stream fusion as in MultiMAE [34], as concatenating RGB and depth tokens increases attention cost proportional to the total token count.

C. Training Objectives

For training, we use Hungarian matching to assign predicted query masks to ground-truth masks [6], [42]. We apply the standard mask-and-class matching cost defined in [6], [7] and omit the class term for embedding supervision. The resulting assignment is reused for all query token-based heads.

For supervision, we use a mask-based segmentation objective and apply the same losses to multi-stage predictions from intermediate query stages, as done in EoMT [9]. This provides earlier training signals to the injected query tokens and stabilizes optimization. For each matched query token, we supervise the predicted mask using Dice loss and a point-sampled binary cross-entropy loss with uncertainty-based point sampling, following Mask2Former [7]. For semantic and panoptic segmentation, we supervise the class logits of matched query tokens using cross-entropy loss. The same strategy is applied for closed-set scene classification.

When enabled, we distill text-aligned visual embeddings using cosine loss and offline teacher targets from Alpha-CLIP [39]. For mask-level embeddings, each matched query token is supervised with a target extracted for the corresponding ground-truth segment using its alpha mask, following DVEFormer [14]. Additionally, for scene-level embeddings, we supervise the embedding with a global target extracted from the full image using the same teacher.

Since only matched queries receive losses, we additionally define a loss term that also targets unmatched queries. Unmatched queries occur when the fixed query token budget K exceeds the number of ground-truth segments in an image and must be discouraged from producing confident outputs. Most mask-based segmentation approaches handle this via a dedicated no-object class in the classification head [6], [7], [9], so that all queries receive a training signal and unmatched queries do not produce confident foreground predictions during inference. To keep unmatched handling consistent across heads–including the embedding head where a no-object target is not well-defined at all–we instead regularize the head outputs (logits or embeddings) for matched query tokens ($i \in \mathcal{M}$) and unmatched query tokens ($i \in \mathcal{U}$) with a loss $\mathcal{L}$. For a given mask-level classification or embedding head, we denote its output vector for the query token i as $\mathbf{v}_i$. Unmatched outputs are encouraged to have a small norm, while matched outputs are constrained to exceed a small minimum norm τ:

$$\mathcal{L} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \|\mathbf{v}_i\|^2 + \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \max(0, \tau - \|\mathbf{v}_i\|)^2$$

For class logits, this primarily discourages high-confidence unmatched predictions, while cross entropy on matched query tokens determines the class decision. For embeddings, we optimize cosine similarity on matched queries using normalized

vectors, so a small lower bound on the pre-normalization norm does not conflict with learning the embedding direction as long as τ is chosen small.

We use this regularization in all our experiments, even if a no-object target would be possible, and thereby show that it serves as an effective alternative across tasks and output heads.

IV. Experiments

In this section, we evaluate IRSAFormer on commonly used RGB-D benchmarks for indoor scene analysis. We conduct experiments on NYUv2 [15], SUN RGB-D [16], and ScanNet V2 [17]. We start by analyzing the semantic segmentation performance under different RGB-D fusion strategies with varying backbone pre-training configurations and model scales on NYUv2 [15]. Based on these findings, the most promising configurations are then evaluated across the datasets for scene classification, semantic segmentation, and panoptic segmentation, as well as text-aligned visual embedding prediction, and compared with the state of the art.

A. Evaluation Metrics

Performance is reported for three closed-set tasks using standard metrics [1], [5], [33]. Scene classification quality is measured by balanced accuracy (bAcc), semantic segmentation quality by mean intersection over union (mIoU), and panoptic segmentation quality by panoptic quality (PQ) [2].

Embedding prediction is evaluated separately, since text-aligned visual embeddings do not directly correspond to a fixed label space and require a mapping from embeddings to classes. Following the evaluation protocol of DVEFormer [14], three complementary strategies are used to obtain closed-set scores from embedding predictions. Text-based retrieval assigns each prediction to the class whose text embedding maximizes cosine similarity. Visual-mean retrieval replaces text prompts by per-class mean embeddings computed from the training set. Finally, linear probing trains a linear classifier on frozen features and provides a strong closed-set reference.

B. Implementation Details

All models were trained and evaluated at an input resolution of 640×480, which is common for indoor robotic perception, and using a patch size of 16. Training employed automatic mixed precision (AMP) to reduce memory consumption and accelerate training. Across all model sizes, query tokens were injected into the last three blocks of the Vision Transformer (ViT) backbone. Query token budgets were kept fixed across tasks and experiments with $K{=}128$ query tokens.

Optimization was performed with AdamW [43] and a batch size of 8. Learning rates were selected by a grid search over $\{1, 2, 3, 4, 5\} \times 10^{-4}$ and scheduled with a polynomial decay. A two-stage warm-up was adopted by freezing the pre-trained backbone for the first 20% of epochs and training only randomly initialized parameters (query tokens and task heads). Afterward, the backbone was unfrozen and fine-tuned jointly with the task-specific parameters for the remaining epochs. Following EoMT [9], mask annealing was applied to stabilize optimization when injecting query tokens to enable standard self-attention at inference time. The loss parameter τ was set to a fixed value of 0.1. Models were trained for 100 epochs on NYUv2, while mixed-dataset training was performed for 25 epochs. Small and Small+ models were trained on single NVIDIA RTX 2080 Ti GPUs with 12 GB VRAM, whereas Base and Large models were trained on single NVIDIA H100 GPUs with 96 GB VRAM.

Following DVEFormer [14], text-aligned teacher targets were generated offline using Alpha-CLIP [39]. Each ground-truth segment was passed as an alpha mask together with the RGB image to the model to extract embeddings. All experiments used the Alpha-CLIP checkpoint based on CLIP-L/14@336 [27] fine-tuned on GRIT-20M. The embedding head therefore predicts $D{=}768$-dimensional text-aligned visual embeddings. For linear probing, an additional linear classifier was trained on top of the frozen model to map embeddings to the closed-set labels.

We provide the network weights and additional training details in our GitHub repository.

C. Depth as Additional Input Modality

Depth provides complementary geometric cues and is known to improve performance on many indoor perception tasks. Therefore, experiments for semantic segmentation on NYUv2 were conducted to compare different RGB-D fusion strategies using ViT-Small backbones with varying pre-training configurations. Fig. 3 presents results for ImageNet-1K [24] (DeiT3-Small-1K [44]) and ImageNet-21K (DeiT3-Small-21K [44]) pre-training as well as larger-scale configurations, including EVA-02-Small (ImageNet-21K distillation) [29] and DINOv3-Small [26] (large-scale self-supervised learning).

For models initialized from the supervised ImageNet checkpoints (DeiT3-Small-1K and DeiT3-Small-21K), incorporating depth consistently improves mIoU for all evaluated RGB-D configurations. The dual-stream setting yields the largest improvements over RGB-only, indicating that a dedicated depth backbone can be beneficial when the backbone pre-training is comparatively limited. Switching to stronger pre-training configurations (EVA-02 and DINOv3) improves segmentation

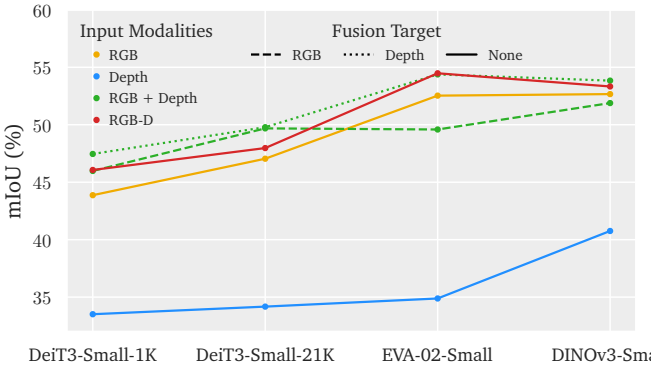

Fig. 3. Semantic segmentation results (mIoU) on NYUv2 (test) for ViT-Small backbones under varying pre-training configurations and RGB-D incorporation strategies. RGB-D denotes early fusion, and RGB + Depth denotes dual-stream fusion. Colors indicate the input modality, and line style indicates the fusion target for dual-stream variants (fusion into RGB or depth, or no fusion).

performance across all modalities. The gap between RGB-only and RGB-D performance narrows, and the difference between early fusion in a single backbone and dual-stream fusion becomes smaller. While dual-stream fusion still tends to yield the best results, its advantage over early fusion is reduced for the strongest pre-training configurations.

Across all pre-training configurations, processing fused features in the depth stream is the most robust choice. Fig. 4 provides qualitative evidence for this observation for ImageNet-21K and DINOv3 pre-trainings, where additional depth yields clearer mask separation under weaker pre-training and fewer mask artifacts under strong pre-training.

For large-scale pre-training in Fig. 3, fusion into the RGB stream can reduce performance, whereas fusion into depth remains stable or improves results. This indicates that fusion into the RGB stream is more sensitive for strongly pre-trained backbones, whereas the depth stream benefits more consistently from fusion. Depth-only performance is consistently below RGB-only across all checkpoints, which is expected given the limited semantic information in depth images. Notably, DINOv3 attains stronger depth-only results than the supervised checkpoints, indicating better utilization of depth cues.

Overall, DINOv3 achieves the best results across fusion settings and is therefore used in subsequent sections.

D. Impact of Model Scale on Performance and Throughput

In this experiment, we analyze the speed-accuracy trade-off of IRSAFormer on an NVIDIA Jetson AGX Orin. Multiple DINOv3 ViT model scales are evaluated for semantic segmentation on NYUv2, comparing input modalities as well as fusion strategies. Fig. 5 summarizes the results.

Scaling from Small to Small+ yields higher mIoU with only a minor reduction in throughput. Moving to the Base backbone further improves semantic segmentation performance while still satisfying our real-time constraint of at least 25 FPS, making it a strong operating point for deployment. The Large backbone achieves the highest overall mIoU but does not meet our real-time requirement and is therefore less suitable.

Adding depth on top of RGB leads to small but consistent improvements in mIoU across all model sizes, except for dual-stream fusion in the RGB backbone. The dual-stream

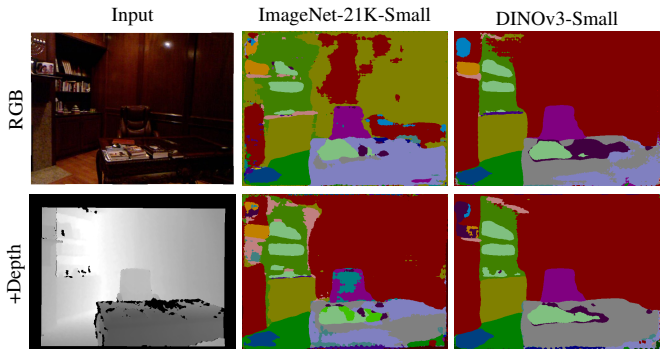

Fig. 4. Qualitative example on NYUv2 (test) comparing RGB-only to dual-stream RGB+Depth (fusion into depth) under ImageNet-21K and DINOv3 pre-training. Semantic colors are chosen as in [1] and are the default colors.

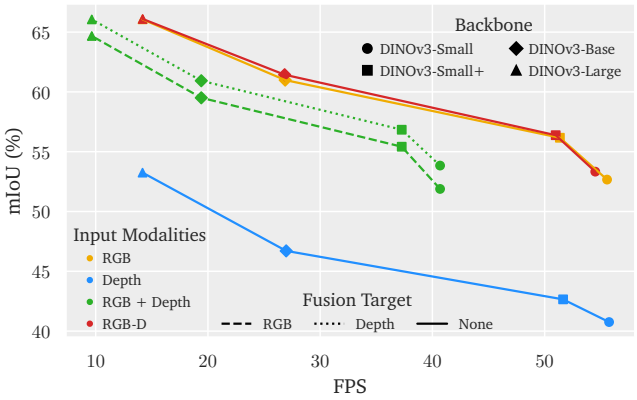

Fig. 5. Semantic segmentation results (mIoU) on NYUv2 (test) over inference throughput (FPS) for different backbone scales and RGB-D integration strategies. RGB-D denotes early fusion, and RGB + Depth denotes dual-stream fusion. Throughput was measured on an NVIDIA Jetson AGX Orin (JetPack 6.2, TensorRT 10.3, Float16 (FP16), 50 W).

design noticeably reduces throughput and rarely offers sufficient performance gains to justify the additional computational overhead. Nevertheless, Fig. 3 suggests that dual-stream fusion can be more beneficial for weaker pre-training configurations, e.g., when large-scale pre-trained checkpoints are unavailable or when experimenting with modified backbones that cannot leverage such initialization.

Overall, RGB-only already performs strongly with DINOv3 backbones, and single-stream RGB-D provides small but consistent improvements at little cost.

E. Robustness to Illumination Changes

Illumination changes are common in mobile robotics, e.g., when the robot moves from dimly lit corridors into brightly lit areas or when strong exposure variations occur due to windows. Such effects can degrade RGB image quality and affect recognition. To study robustness to illumination changes in a controlled setting, robustness is evaluated by applying gamma correction to the RGB input at inference time while keeping the depth input unchanged. Fig. 6 reports results for mIoU on NYUv2 for the selected DINOv3-Base backbone.

Performance is highest at the original input and degrades with increasing levels of brightening or darkening. Adding

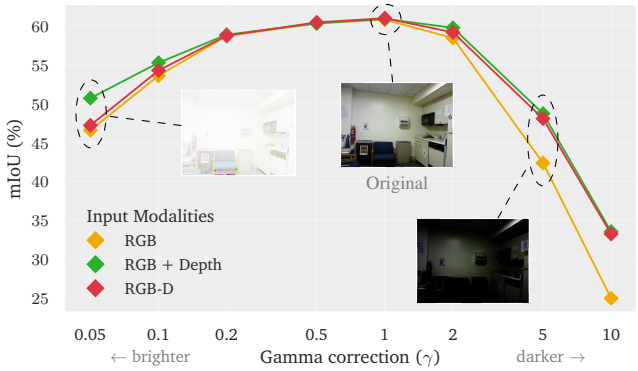

Fig. 6. Robustness to illumination changes on NYUv2 (test) using the DINOv3-Base backbone. Gamma correction is applied to the RGB input at inference time while keeping the depth input unchanged.

depth consistently improves robustness compared to RGB-only in such settings. Under strong overexposure, the dual-stream design with fusion into the depth stream retains higher accuracy than early-fusion RGB-D, whereas under strong underexposure both RGB-D variants behave similarly. This suggests that explicit depth processing can be particularly beneficial when RGB appearance cues are saturated, while early fusion already captures most of the benefit in darker scenarios.

F. State-of-the-Art Comparison and Embedding Estimation

Tab. I presents a comparison of IRSAFormer with state-of-the-art indoor scene analysis approaches on NYUv2, SUN RGB-D, and ScanNet V2 for scene classification, semantic segmentation, and panoptic segmentation. For IRSAFormer, we report results for both RGB-only and single-stream early-fusion RGB-D variants. Since training settings differ across the approaches, Tab. I highlights dataset-specific fine-tuning and also denotes whether additional training data beyond an ImageNet pre-training and the target benchmark was used. Specifically, for practical robotics deployment, we train a single generalist IRSAFormer per model size by mixing indoor datasets following [37] and evaluate it directly on each benchmark without additional dataset-specific fine-tuning.

Across all datasets and all three tasks, IRSAFormer achieves consistently strong closed-set performance. For scene classification, IRSAFormer with smaller backbones matches previous approaches and improves further with larger backbones. For semantic segmentation, the DINOv3-Small+ backbone already outperforms most competing approaches, including several larger backbones, and switching to DINOv3-Base further improves results while maintaining real-time throughput on an NVIDIA Jetson AGX Orin. Scaling to DINOv3-Large yields the best overall mIoU values and surpasses prior strong baselines such as OmniVec2 [47]. For panoptic segmentation, IRSAFormer also reaches competitive PQ on all three datasets. The smaller variants are on par with other RGB-D approaches, while larger backbones further improve PQ at the expected cost of reduced throughput.

Comparing input modalities, for all DINOv3 backbone initializations, RGB-only and single-stream RGB-D results are close, and the preferred modality can depend on the dataset and model scale. However, standard benchmarks mainly reflect the recording conditions of the underlying datasets, and samples with extreme exposure are relatively rare. Therefore, the reported metrics may not fully reflect robustness under conditions frequently encountered by real-world mobile robots.

Beyond closed-set prediction, IRSAFormer can optionally predict text-aligned visual embeddings for each query token and a scene-level embedding, enabling retrieval-style evaluation as in DVEFormer [14]. The results in Tab. I reveal that token-level embeddings of IRSAFormer are more difficult to align to text than dense per-pixel embeddings, which reduces text-based and visual-mean retrieval performance compared to DVEFormer. However, at the same time, linear probing shows that the learned token embeddings still encode meaningful semantic information, even when direct text alignment is weaker. Scene-level embeddings are distilled more reliably and yield competitive retrieval-based scene prediction, with visual-mean retrieval often approaching closed-set classification performance. These findings suggest that stronger text-alignment objectives or more specialized projection heads, as explored in recent work on vision-language alignment (e.g., DINO-style

TABLE I

State-of-the-art comparison on NYUv2, SUN RGB-D, and ScanNet V2 for scene classification (bAcc), semantic segmentation (mIoU), and panoptic segmentation (PQ), as well as throughput measured on an NVIDIA Jetson AGX Orin (JetPack 6.2, TensorRT 10.3, FP16, 50 W) at 640×480. IRSAFormer results are presented for RGB-only and early-fusion RGB-D. For the embedding variant (Embeddings), linear probing results are shown in black, and text-based / visual mean in gray. Legend: * – results with test-time augmentation; ▽ – expected to be lower; and ** – deviating input resolution of 1024×768.

	Model	Backbone		Extra Data	NYUv2 Test			SUN RGB-D Test			ScanNet V2 (20 Classes) Valid			ScanNet V2 (20 Classes) Test		
					bAcc↑	mIoU↑	PQ↑	bAcc↑	mIoU↑	PQ↑	bAcc↑	mIoU↑	PQ↑	mIoU↑	FPS↑
with dataset-specific fine-tuning	EMSAFormer [33]	SwinV2-T-128			80.02	50.51	43.41	64.50	48.82	51.70	49.73	64.75	51.18	56.4	39.10
	DFormerv2 [45]	DFormerv2-B		✓	-	57.7	-	-	52.8	-	-	-	-	-	18.25
		DFormerv2-L		✓	-	58.4	-	-	53.3	-	-	-	-	-	13.24
	CMX [4]	2x MiT-B2			-	54.1	-	-	49.7*	-	-	61.3	-	-	33.37
		2x MiT-B5			-	56.8	-	-	52.4*	-	-	-	-	-	- ▽
	CMNext [46]	2x MiT-B4			-	56.9	-	-	51.9	-	-	-	-	-	14.84
	Omnivore [32]	Swin-B		✓	-	54.0	-	-	-	-	-	-	-	-	- ▽
	OmniVec2 [47]	BERT		✓	-	63.6	-	-	-	-	-	-	-	-	- ▽
	MultiMAE [34]	ViT-B		✓	-	56.0	-	-	51.1	-	-	-	-	-	- ▽
	Vanishing Depth [35]	2x ViT-B		✓	-	55.07	-	-	56.05	-	-	-	-	-	- ▽
without dataset-specific fine-tuning	EMSANet [1], [37]	2x ResNet34-NBt1D		✓	70.29	56.55	48.47	58.56	49.31	50.91	49.01	66.93	53.92	60.0	38.57**
	DVEFormer [14]	SwinV2-T-128		✓	-	57.07	-	-	50.99	-	-	67.72	-	62.6	26.30
				✓		44.07 / 50.31			44.56 / 46.25			50.16 / 56.57			
	IRSAFormer (Ours)	DINOv3-Small+	RGB	✓	80.61	58.26	47.86	65.65	54.03	52.40	48.00	63.96	52.53	-	51.36
			+D	✓	79.73	57.93	48.23	66.04	52.34	53.47	51.37	65.96	54.52	-	50.98
	IRSAFormer (Ours)	DINOv3-Base	RGB	✓	82.21	63.20	49.90	68.46	56.20	55.97	53.37	68.40	55.01	-	26.88
			+D	✓	80.94	62.67	50.59	65.93	56.30	52.11	52.62	68.99	54.82	64.0	26.83
	IRSAFormer (Ours) (Embeddings)	DINOv3-Base	+D	✓	84.57	60.93	49.26	69.98	46.70	45.79	55.03	66.61	49.11	-	26.79
				✓	72.40 / 82.03	41.06 / 49.20	36.83 / 42.31	60.45 / 65.49	38.71 / 39.63	35.42 / 37.76	50.41 / 54.63	46.85 / 53.99	37.77 / 43.37		
	IRSAFormer (Ours)	DINOv3-Large	RGB	✓	85.43	67.16	54.56	70.97	57.69	59.58	56.74	72.72	60.84	-	14.17
			+D	✓	83.56	67.62	56.14	70.88	57.63	58.56	56.80	70.62	58.83	-	14.19

text alignment [40]), may be required to improve token-level embeddings for text alignment. Nevertheless, predicting embeddings per token remains attractive for robotic deployment, since it is significantly more memory-efficient than dense per-pixel embeddings while retaining the same flexibility.

V. Conclusion

In this paper, we have presented IRSAFormer, an efficient encoder-centric Transformer-based approach for indoor robotic scene analysis. IRSAFormer injects query tokens into a largely unmodified ViT backbone and uses lightweight task heads to decode tokens for scene classification, semantic segmentation, and panoptic segmentation. A central focus was the role of depth. Experiments across common indoor datasets have proven that depth provides clear gains for limited-scale pre-training, whereas large-scale pre-training substantially narrows the gap between RGB and RGB-D, making RGB-only models highly competitive. In addition, we showed that IRSAFormer can predict text-aligned visual embeddings as a more flexible output representation for language-driven downstream applications, although current results are more reliable at scene level than at query token level. Overall, IRSAFormer achieves state-of-the-art performance while reaching 26.8 FPS on an NVIDIA Jetson AGX Orin with a ViT-Base backbone, allowing real-time downstream robotic applications.

References

[1] D. Seichter, S. B. Fischedick, M. Köhler, and H.-M. Gross, "Efficient Multi-Task RGB-D Scene Analysis for Indoor Environments," in *Proc. of IJCNN*, 2022.

[2] A. Kirillov, K. He, R. Girshick *et al.*, "Panoptic Segmentation," in *Proc. of CVPR*, 2019, pp. 9404–9413.

[3] B. Cheng, M. D. Collins, Y. Zhu *et al.*, "Panoptic-DeepLab: A Simple, Strong, and Fast Baseline for Bottom-Up Panoptic Segmentation," in *Proc. of CVPR*, 2020, pp. 12 475–12 485.

[4] J. Zhang, H. Liu, K. Yang *et al.*, "CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers," *ITS*, vol. 24, no. 12, pp. 14 679–14 694, 2023.

[5] B. Yin, X. Zhang, Z.-Y. Li *et al.*, "DFormer: Rethinking RGBD representation learning for semantic segmentation," in *Proc. of ICLR*, 2024.

[6] B. Cheng, A. G. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. of NeurIPS*, 2021.

[7] B. Cheng, I. Misra, A. G. Schwing *et al.*, "Masked-attention mask transformer for universal image segmentation," in *Proc. of CVPR*, 2022.

[8] Z. Li, W. Wang, E. Xie *et al.*, "Panoptic SegFormer: Delving deeper into panoptic segmentation with transformers," in *Proc. of CVPR*, 2022, pp. 1280–1289.

[9] T. Kerssies, N. Cavagnero, A. Hermans *et al.*, "Your ViT is Secretly an Image Segmentation Model," in *Proc. of CVPR*, 2025.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *Proc. of ICLR*, 2021.

[11] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating Depth into Semantic Segmentation via Fusion-based CNN Architecture," in *Proc. of ACCV*, 2016, pp. 213–228.

[12] D. Seichter, M. Köhler, B. Lewandowski *et al.*, "Efficient RGB-D Semantic Segmentation for Indoor Scene Analysis," in *Proc. of ICRA*, 2021.

[13] C. Zhou, C. C. Loy, and B. Dai, "Extract Free Dense Labels from CLIP," in *Proc. of ECCV*, 2022.

[14] S. Fischedick, D. Seichter, B. Stephan *et al.*, "Efficient Prediction of Dense Visual Embeddings via Distillation and RGB-D Transformers," in *Proc. of IROS*, 2025.

[15] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor Segmentation and Support Inference from RGBD Images," in *Proc. of ECCV*, 2012.

[16] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite," in *Proc. of CVPR*, 2015.

[17] A. Dai, A. X. Chang, M. Savva *et al.*, "ScanNet: Richly-annotated 3D Reconstructions of Indoor Scenes," in *Proc. of CVPR*, 2017.

[18] L.-C. Chen, Y. Zhu, G. Papandreou *et al.*, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Proc. of ECCV*, 2018, pp. 801–818.

[19] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *Proc. of ICCV*, 2017.

[20] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic Feature Pyramid Networks," in *Proc. of CVPR*, 2019, pp. 6399–6408.

[21] Y. Xiong, R. Liao, H. Zhao *et al.*, "UPSNet: A Unified Panoptic Segmentation Network," in *Proc. of CVPR*, 2019, pp. 8818–8826.

[22] J. Šarić, M. Oršić, and S. Segvic, "Panoptic swiftnet: Pyramidal fusion for real-time panoptic segmentation," *arXiv preprint arXiv:2203.07908*, 2022.

[23] H. Wang, Y. Zhu, H. Adam *et al.*, "MaX-DeepLab: End-to-End Panoptic Segmentation with Mask Transformers," in *Proc. of CVPR*, 2021, pp. 5463–5474.

[24] O. Russakovsky, W. Dong, R. Socher *et al.*, "ImageNet Large Scale Visual Recognition Challenge," in *IJCV*, 2015, pp. 211–252.

[25] M. Oquab, T. Darcet, T. Moutakanni *et al.*, "DINOv2: Learning Robust Visual Features without Supervision," *arXiv:2304.07193*, 2023.

[26] O. Siméoni, H. V. Vo, M. Seitzer *et al.*, "DINOv3," *arXiv preprint arXiv:2508.10104*, 2025.

[27] A. Radford, J. W. Kim, C. Hallacy *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. of ICML*, 2021, pp. 8748–8763.

[28] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *Proc. of NeurIPS (Workshop)*, 2015.

[29] Y. Fang, Q. Sun, X. Wang *et al.*, "EVA-02: A visual representation for neon genesis," *IVC*, vol. 149, p. 105171, 2024.

[30] L. Yang, B. Kang, Z. Huang *et al.*, "Depth anything v2," *Proc. of NeurIPS*, vol. 37, pp. 21 875–21 911, 2025.

[31] M. Cao, H. Leng, D. Lischinski, and Y. Li, "ShapeConv: Shape-aware Convolutional Layer for Indoor RGB-D Semantic Segmentation," in *Proc. of ICCV*, 2021.

[32] R. Girdhar, M. Singh, N. Ravi *et al.*, "Omnivore: A Single Model for Many Visual Modalities," in *Proc. of CVPR*, 2022.

[33] S. Fischedick, D. Seichter, R. Schmidt *et al.*, "Efficient Multi-Task Scene Analysis with RGB-D Transformers," in *Proc. of IJCNN*, 2023.

[34] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "MultiMAE: Multi-modal multi-task masked autoencoders," in *Proc. of ECCV*, 2022.

[35] P. Koch, J. Krüger, A. Chowdhury, and O. Heimann, "Vanishing depth: A depth adapter with positional depth encoding for generalized image encoders," *arXiv preprint arXiv:2503.19947*, 2025.

[36] D. Seichter, P. Langer, T. Wengefeld *et al.*, "Efficient and Robust Semantic Mapping for Indoor Environments," in *Proc. of ICRA*, 2022.

[37] D. Seichter, B. Stephan, S. B. Fischedick *et al.*, "PanopticNDT: Efficient and Robust Panoptic Mapping," in *Proc. of IROS*, 2023, pp. 7233–7240.

[38] S. Cho, H. Shin, S. Hong *et al.*, "Cat-seg: Cost aggregation for open-vocabulary semantic segmentation," in *Proc. of CVPR*, 2024, pp. 4113–4123.

[39] Z. Sun, Y. Fang, T. Wu *et al.*, "Alpha-CLIP: A CLIP model focusing on wherever you want," in *Proc. of CVPR*, 2024, pp. 13 019–13 029.

[40] C. Jose, T. Moutakanni, D. Kang *et al.*, "DINOv2 meets text: A unified framework for image-and pixel-level vision-language alignment," in *Proc. of CVPR*, 2025, pp. 24 905–24 916.

[41] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. of CVPR*, 2018, pp. 7132–7141.

[42] N. Carion, F. Massa, G. Synnaeve *et al.*, "End-to-end object detection with transformers," in *Proc. of ECCV*. Springer, 2020, pp. 213–229.

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. of ICLR*, 2019.

[44] H. Touvron, M. Cord, and H. Jegou, "DeiT III: Revenge of the ViT," *Proc. of ECCV*, 2022.

[45] B.-W. Yin, J.-L. Cao, M.-M. Cheng, and Q. Hou, "DFormerv2: Geometry self-attention for RGBD semantic segmentation," in *Proc. of CVPR*, 2025.

[46] J. Zhang, R. Liu, H. Shi *et al.*, "Delivering arbitrary-modal semantic segmentation," in *Proc. of CVPR*, 2023.

[47] S. Srivastava and G. Sharma, "Omnivec2-a novel transformer based network for large scale multimodal and multitask learning," in *Proc. of CVPR*, 2024.