AdapToPASS: Ambiguity-aware Adaptive Spherical Transformer for PAnoramic Semantic Segmentation

Soumyaratna Debnath[1] **Weiming Zhang** [2] **Shriram Damodaran**[1]
Dingwen Xiao[2] **Lin Wang**[1,†]

[1]EmPACT Lab, School of EEE, Nanyang Technological University, Singapore
[2]Hong Kong University of Science and Technology (Guangzhou)

soumyara004@e.ntu.edu.sg, wzhang915@connect.hkust-gz.edu.cn,
shriram003@e.ntu.edu.sg, dxiaoaf@connect.hkust-gz.edu.cn, linwang@ntu.edu.sg

[†]Corresponding author

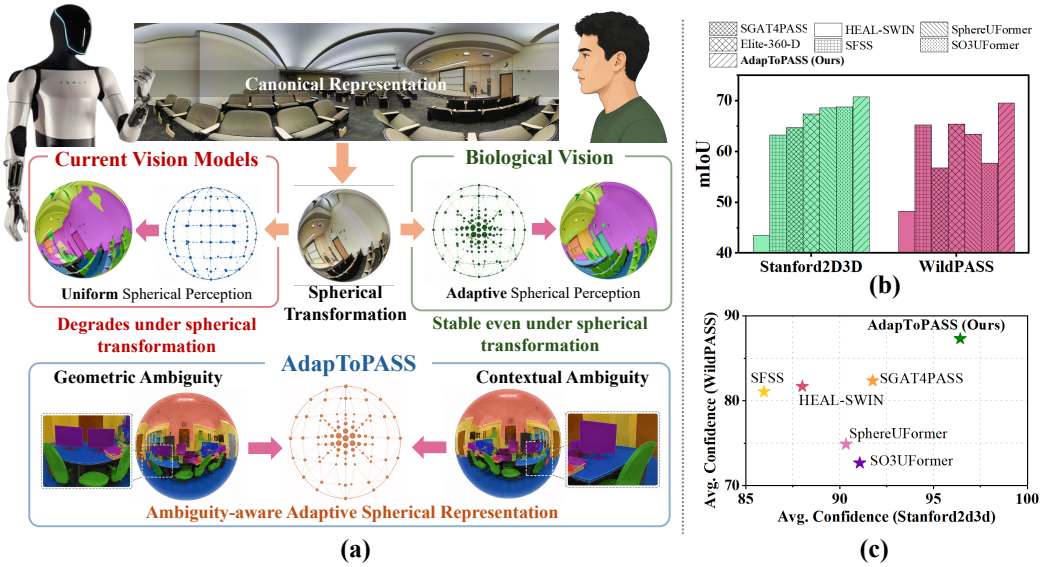


Figure 1: **(a)** Spherical perception in current methods degrades under spherical transformations due to geometric and contextual ambiguity, while biological vision remains stable via adaptive perception; AdapToPASS models this adaptivity. **(b)** Improved segmentation performance of AdapToPASS on Stanford2D3D and WildPASS. **(c)** AdapToPASS exhibits higher and more consistent confidence, indicating stronger geometric and contextual robustness. The project is publicly available at https://empactlab.github.io/AdapToPASS-NeurIPS-2026/.

Abstract

Spherical Transformers have emerged as a promising framework for panoramic semantic segmentation (PASS) by operating directly on spherical geometry and alleviating projection-induced distortions. However, existing architectures rely on assumptions of canonical spherical structure and stable viewpoints, which are frequently violated in real-world 360° imagery due to unconstrained camera motion, introducing significant contextual and geometric ambiguity. Consequently, they lack adaptive mechanisms to model such ambiguity, limiting robustness to unseen spherical transformations. In contrast, biological perception is inherently ambiguity-aware: rather than estimating uncertainty probabilistically, it adapts to fluctuations in cue reliability caused by geometric and contextual variations,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

enabling stable interpretation under complex visual transformations. Motivated by these observations, we first present a **systematic analysis** of existing PASS architectures under various unseen spherical transformations. Following this, we introduce **AdapToPASS**, a **novel**, **bio-inspired** Spherical Transformer that adaptively models contextual and geometric ambiguities for robust panoramic semantic segmentation. At the core of AdapToPASS are the **Adaptive Spherical Attention (AdaSpA)** blocks that dynamically modulate attention according to local contextual ambiguity, mimicking the adaptive, context-driven perception of biological vision. To address geometric ambiguities, AdapToPASS employs **Bifocal Spherical Representation** that reconciles the trade-off between field of view and spatial resolution; along with *boundary supervision* to emulate the boundary-sensitive nature of biological vision. We evaluate AdapToPASS in both indoor and outdoor semantic segmentation, where it consistently outperforms prior state-of-the-art methods. We further validate AdapToPASS under unseen spherical transformations, where it demonstrates strong robustness and surpasses the next-best method by **+13.38%** relative improvement in mIoU on **Stanford2D3D** and **+18.77%** on **WildPASS**. Additionally, we introduce a lightweight variant, **AdapToPASS-Swift**, with fewer than **2M parameters**, which surpasses compact baselines while retaining robustness to spherical transformations.

1 Introduction

Panoramic semantic segmentation plays a crucial role in perception for applications such as robotic navigation and virtual reality, where understanding the full 360° surrounding environment is essential for decision-making [1, 2, 3, 4, 5]. Although semantic segmentation has achieved remarkable success in perspective images with Convolutional Neural Networks (CNNs), encoder-decoder architectures, and Transformer-based models [6, 7, 8, 9, 10], these approaches are inherently designed for perspective 2D images and fail to capture the unique geometric properties of spherical data [11]. A common workaround is to project spherical images into the equirectangular projection (ERP) for compatibility with standard 2D architectures. However, ERP introduces severe distortions, breaks the left-right continuity at image boundaries, and spans an extremely large field-of-view (FoV). This limits the effectiveness of the 2D architectures [12, 13, 14, 15, 16]. To overcome this limitation, several spherical Transformer backbones have been introduced to operate directly on the sphere [17, 18]. Despite their strong performance, these models rely on assumptions of canonical spherical structure and stable viewpoints. In practice, however, omnidirectional cameras on handheld devices, and mobile robot platforms undergo unconstrained motion, introducing arbitrary rotations, scale changes, and viewpoint shifts that violate these assumptions and substantially degrade segmentation accuracy [19, 20].

In this work, we first conduct a systematic analysis of commonly used spherical Transformer backbones under a diverse set of spherical transformations, which revealed clear performance degradation patterns (see Fig. 2). More critically, our analysis reveals **two** fundamental limitations in these architectures: **(1) Contextual ambiguity.** Existing methods rely on uniform feature aggregation regardless of scene complexity, failing to adapt when local regions exhibit ambiguous or overlapping class boundaries (Fig 1a). **(2) Geometric ambiguity.** They treat all spatial locations equally, without the varying degrees of distortion and structural irregularity introduced by spherical transformations.

From a neuroscience perspective, biological vision is inherently adaptive, multi-scale, and ambiguity-aware. Rather than processing the entire visual field uniformly, the human visual system dynamically adjusts contextual integration according to local ambiguity, and expands receptive context in uncertain regions while preserving fine detail in reliable ones [21, 22, 23, 24, 25]. It further balances a trade-off between FoV and spatial resolution to resolve ambiguity across different spatial scales, which enables both global scene understanding and high-acuity local perception [26, 27, 28]. Moreover, biological perception is boundary-aware, as it helps disambiguate object contours and region transitions even under appearance and viewpoint changes [29]. All these characteristics of biological vision contribute to their remarkable robustness to unseen visual transformations, including changes in orientation, scale, and viewpoint [30, 31, 32, 33] (Fig 1a).

These observations raises an interesting question: ***How can ambiguity-aware perception in human vision be efficiently and adaptively integrated into Spherical Transformers to achieve robust***

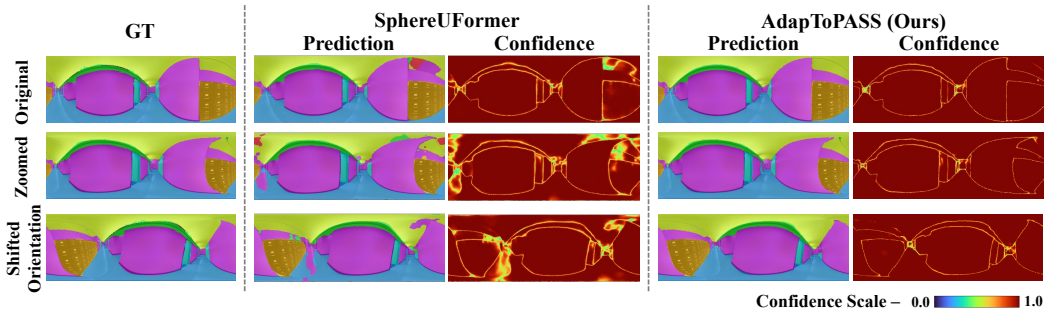


Figure 2: **Qualitative comparison of AdapToPASS and SphereUFormer on Stanford2D3D under unseen spherical transformations.** We show segmentation predictions and confidence maps, where **confidence is defined as the inverse of ambiguity** $\text{Conf} = 1 - \text{Amb}$ and visualized using the maximum softmax probability. SphereUFormer shows a clear confidence drop under spherical transformations, whereas AdapToPASS maintains higher confidence and sharper semantic boundaries.

panoramic semantic segmentation? To this end, we propose **AdapToPASS**, a bio-inspired panoramic semantic segmentation framework that models: (**I**) contextual ambiguity via **Adaptive Spherical Attention (AdaSpA)** blocks, which adaptively modulate attention according to local contextual ambiguity; and (**II**) geometric ambiguity via *Bifocal Spherical Representation* with dual pathways, an **Acuity Stream** focusing on high-resolution detail and a **Lateral Stream** focusing on wide FoV context, effectively balancing the FoV vs. resolution trade-off. Additionally, we incorporate *SDF-based boundary supervision* to explicitly capture boundary ambiguity and improve localization.

We evaluate AdapToPASS across both indoor and outdoor panoramic semantic segmentation benchmarks. We further assess generalization under **unseen spherical transformations** across varying difficulty levels. AdapToPASS it surpasses the previous best-performing method by **+13.38%** *relative improvement in mIoU* on **Stanford2D3D** [34] and **+18.77%** on **WildPASS** [35, 11].

To summarize, the major contributions of this paper are four-fold:

- Inspired by neuroscience-grounded principles, we identify a fundamental limitation of existing panoramic segmentation models and conduct a systematic analysis of their generalization under unseen spherical transformations.
- We propose *AdaSpA* blocks that adaptively modulate attention according to local contextual ambiguity, serving as a computational analogue of uncertainty-driven semantic processing in biological vision (Sec 3.1).
- We introduce the *Bifocal Spherical Representation* framework, inspired by the biological trade-off between spatial acuity and field of view (Sec. 3.2), which serves as a computational analogue for addressing geometric ambiguities.
- We validate our method on both indoor and outdoor panoramic segmentation benchmarks, and evaluate its generalization under unseen spherical transformations (Sec. 4). We further introduce a compact 2M-parameter variant, showing that ambiguity-aware spherical perception can be preserved under strict efficiency constraints (Sec. 4.2).

2 Related Works

Recently, bio-inspired mechanisms have gained widespread adoption across diverse domains including art, robotics and material design [36, 37, 38, 39].

Panoramic Semantic Segmentation. For 360° imagery, equirectangular projection (ERP) remains the most widely used representation due to its simplicity and compatibility with standard vision architectures. However, ERP introduces severe latitude-dependent distortions, especially near the poles, which compromise geometric consistency and semantic reasoning [12, 40, 13, 41, 14, 42]. Distortion-aware Transformers such as Trans4PASS [43] and PanoFormer [44] mitigate ERP distortions through deformable or sphere-aware tokenization, but remain tied to planar projections. A more natural alternative is to operate directly on the sphere, to avoid artifacts caused by planar parameterization. Foundational efforts such as spherical harmonic networks [13, 45] and gauge-equivariant CNNs [46] established principled frameworks for spherical representation learning. Building on this direction, recent panoramic segmentation methods increasingly incorporate geometry-aware or

native spherical representations. HEAL-SWIN [17] combines the equal-area HEALPix grid with hierarchical attention; SFSS [47] explores multimodal spherical segmentation using RGB, depth, and normal cues. Similarly, Elite360D [48] introduces icosahedron-based bi-projection fusion for efficient panoramic dense perception, and later extended to multi-task panoramic perception and segmentation. Most recently, SphereUFormer [18] proposed a U-shaped Transformer on the icosphere with localized spherical self-attention and geometry-consistent multi-scale processing, achieving state-of-the-art performance in panoramic semantic segmentation. *Despite these advances, current panoramic segmentation methods largely assume a stable canonical spherical layout during training and evaluation. Their robustness to unseen spherical transformations, including rotations, scale changes, orientation shifts, and viewpoint shifts, remains insufficiently studied.*

Transformation Robustness in Panoramic Segmentation. Generalization beyond canonical spherical layouts is critical for real-world panoramic perception. Most panoramic segmentation methods implicitly assume a canonical camera orientation [20]. This assumption is rarely satisfied in real-world 360° imagery, where camera motion and viewpoint changes can induce substantial spherical distortions. Recent efforts have begun to address this fragility via transformation-aware modeling. For example, prior works have explored Möbius transformations for data augmentation [49, 50], while PanDA [51] uses them to enforce consistency under geometric deformations in panoramic depth estimation. Möbius-equivariant CNNs [52] further provide a theoretical foundation for conformal-equivariant spherical CNNs. However, these works do not explicitly study robustness to diverse spherical transformations in panoramic segmentation. Within panoramic segmentation, existing efforts primarily focus only on SO(3) rotations or rely on transformation-specific augmentation. SGAT4PASS [19] improves robustness to 3D disturbances through spherical geometry-aware modeling, while the concurrent SO3UFormer [20] enhances rotation robustness using gauge-aware relative positional mechanisms and data augmentation. More recently, Unified Spherical Frontend [53] achieves cross-camera rotation equivariance by construction through distance-only convolution kernels on icosahedral grids. *However, these methods mainly improve geometric consistency under specific perturbations, rather than providing an adaptive mechanism for broader transformation-induced ambiguity.*

Geometric and Context Modeling for Segmentation. Panoramic semantic segmentation is inherently ambiguous: wide-field imagery introduces unreliable long-range cues, spherical distortions, viewpoint changes, and unstable semantic transitions, leading to contextual, geometric, and boundary ambiguity. In particular, Dinh et al. [54] show that contextual ambiguity in monocular 360° imagery limits uniform feature aggregation, highlighting the need for adaptive context modeling. Existing panoramic methods incorporate structured contextual aggregation [11, 55] or identify semantic/geometric ambiguity in distortion-heavy regions [56], while geometric-aware models mitigate projection-induced distortions through sphere-aware tokenization or spherical representations [44, 17]. Boundary-aware segmentation losses further show that explicitly modeling semantic transitions can reduce diffuse predictions near object boundaries [57]. Beyond segmentation, adaptive models such as HAFA [58] and AMContrast3D [59] demonstrate the value of uncertainty-guided feature modulation. *However, these efforts do not provide a unified mechanism for jointly resolving contextual, geometric, and boundary ambiguity within a native segmentation architecture.*

Takeaway. To our knowledge, this is *the first* unified architectures to jointly model contextual, geometric, and boundary ambiguity for panoramic semantic segmentation. AdapToPASS introduces an ambiguity-aware Spherical Transformer that adaptively aggregates context, handles spherical geometric distortion, and predicts sharper semantic boundaries directly on the sphere.

3 Methodology

Overview. An overview of AdapToPASS is illustrated in Fig. 3. Given an omnidirectional RGB image, we first projects it onto an icosphere, obtaining a spherical signal $\mathbf{x} \in \mathbb{R}^{N\times 3}$ over N icosphere vertices. The model produces per-node semantic logits $\hat{\mathbf{y}} \in \mathbb{R}^{N\times K}$ and per-class signed distance field (SDF) predictions $\hat{\mathbf{s}} \in \mathbb{R}^{N\times K}$, where K is the number of semantic classes. Additional implementation details are provided in Appendix A.1.

Conceptually, we formulate AdapToPASS from an energy-minimization perspective:

$$(\hat{\mathbf{Y}}, \hat{\mathbf{S}}) = \arg\min_{\mathbf{Y},\mathbf{S}} \mathcal{E}_{\text{sem}}(\mathbf{Y}; \mathbf{X}) + \lambda_c \, \mathcal{E}_{\text{ctx}}(\mathbf{Y}; \mathbf{c}, \mathcal{G}) + \lambda_g \, \mathcal{E}_{\text{geo}}(\mathbf{Y}, \mathbf{S}; \mathbf{D}^{\partial}, \mathcal{G}) \tag{1}$$

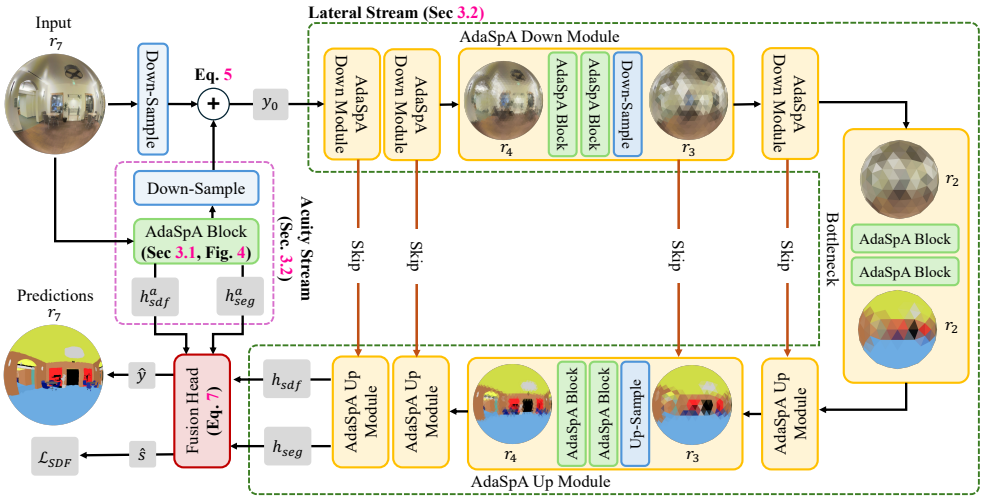


Figure 3: Overview of the AdapToPASS framework. *Zoom in for a better view.*

This formulation estimates semantic predictions $\hat{\mathbf{Y}}$ and SDF predictions $\hat{\mathbf{S}}$ by minimizing three complementary energies: semantic prediction $\mathcal{E}_{\text{sem}}$; contextual ambiguity $\mathcal{E}_{\text{ctx}}$ modeled by AdaSpA through the learned ambiguity signal c_i and the induced contextual-ambiguity-aware geodesic bias $b^h_{ij}(c_i)$ (Sec. 3.1); and geometric ambiguity $\mathcal{E}_{\text{geo}}$ modeled by the Bifocal Spherical Representation Framework (Sec. 3.2). Here, $\mathbf{c}$ controls adaptive context aggregation, while $\mathbf{D}^{\partial}$ provides geometric boundary guidance on the icosphere graph $\mathcal{G}$.

3.1 Adaptive Spherical Attention (AdaSpA) Blocks

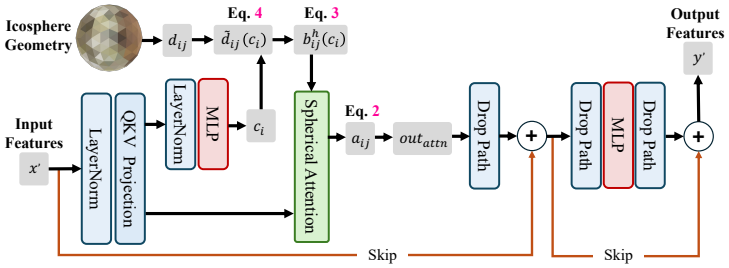


Figure 4: Overview of an AdaSpA Block.

Insight. Standard spherical attention mechanisms aggregate context uniformly across the sphere, assigning equal weight to all neighbors regardless of their local contextual ambiguity [18, 17]. When sensory evidence is ambiguous, human perception relies more on context and priors; when reliable, it follows the local signal. We propose Adaptive Spherical Attention (AdaSpA), which dynamically modulates contextual integration by adapting each token's attention to a learned local contextual ambiguity signal. The overview of the block is illustrated in Fig. 4.

Spherical Attention. At each icosphere rank r, an AdaSpA block performs local self-attention over the w-hop neighborhood $\mathcal{N}_w(i)$ of each node v_i. Given token features $\mathbf{x} \in \mathbb{R}^{N\times d}$, queries, keys, and values are computed from linear projections of neighboring tokens. For head h, the attention logit between node v_i and its neighbor v_j is defined as

$$a^h_{ij} = \tau_h \hat{\mathbf{Q}}^h_i \hat{\mathbf{K}}^{h\top}_j + b^h_{ij}(c_i), \tag{2}$$

where $\hat{\mathbf{Q}}$ and $\hat{\mathbf{K}}$ are ℓ_2-normalized queries and keys, τ_h is a learnable head-specific scale, and $b^h_{ij}(c_i)$ is an ambiguity-aware geodesic bias. This formulation retains standard local self-attention while injecting spherical geometry directly into the attention logits.

Geodesic Bias. To encode the intrinsic geometry of the sphere, we compute the geodesic distance between neighboring icosphere nodes: $d_{ij} = \arccos(\mathbf{n}_i^\top \mathbf{n}_j)$, where $\mathbf{n}_i$ and $\mathbf{n}_j$ are unit sphere normals. Rather than relying only on learned relative position embeddings, we map this distance to a head-specific attention bias:

$$b^h_{ij} = \lambda_{\text{geo}}\, \phi_h(d_{ij}), \tag{3}$$

where $\phi_h(\cdot)$ is a lightweight MLP producing the bias for head h, and λ_{geo} is a learnable scale. This encourages attention to respect spherical proximity while remaining learnable.

Contextual-Ambiguity-Aware Geodesic Bias. A fixed geodesic bias applies the same spatial preference to all tokens, regardless of local semantic difficulty. However, ambiguous regions often require broader contextual evidence, whereas confident regions can rely more strongly on local structure. AdaSpA therefore predicts a contextual ambiguity score $c_i \in [0, 1]$ for each query token and uses it to rescale the geodesic distance:

$$\tilde{d}_{ij}(c_i) = \text{clamp}\left(\frac{d_{ij}}{s_{\min} + (s_{\max} - s_{\min})c_i + \varepsilon}, 0, \pi\right). \tag{4}$$

The modulated distance $\tilde{d}_{ij}(c_i)$ replaces d_{ij} in Eq. (3), yielding the ambiguity-aware bias $b^h_{ij}(c_i)$ used in Eq. (2). When c_i is high, the effective distance profile is compressed, encouraging broader contextual aggregation; when c_i is low, the bias remains sharper and more local. Each AdaSpA block follows a standard pre-normalization attention-FFN structure.

3.2 Bifocal Spherical Representation Framework

Insight. Conventional spherical segmentation models generally follow a single hierarchical pathway, where expanding the receptive field progressively reduces spatial resolution [18, 20]. This creates a fundamental tension between acuity and field of view. Inspired by biological vision, where foveated eyes prioritize high spatial acuity while laterally placed eyes favor wide-field contextual awareness, AdapToPASS introduces a *Bifocal Spherical Representation Framework* with two specialized streams. The **Acuity Stream** focus on the full-resolution icosphere (r=7) to preserve local geometry and boundary detail, and the **Lateral Stream** focus on a downsampled icosphere (r'=r−1=6) to aggregate broader semantic context.

Acuity Stream. The Acuity Stream preserves fine-grained spherical structure that may be weakened by hierarchical downsampling in the Lateral Stream. Given the full-resolution icosphere input $\mathbf{x} \in \mathbb{R}^{N\times 3}$, we first project it into a lightweight feature space and process it with an AdaSpA block, which produce full-resolution acuity features $\mathbf{F}_{\text{acuity}} \in \mathbb{R}^{N\times d_a}$. These features are injected into the Lateral Stream through a gated projection:

$$\mathbf{y}_0 = \mathbf{E}(\mathbf{x}) + \sigma(g_{\text{gate}})\,\Pi(\mathbf{F}_{\text{acuity}}), \tag{5}$$

where $\mathbf{E}(\cdot)$ denotes the Lateral input projection and $\Pi(\cdot)$ downsamples and projects the acuity features to the Lateral resolution. To retain full-resolution supervision, the Acuity Stream also produces auxiliary segmentation and SDF predictions.

Lateral Stream. The Lateral Stream is designed to capture broad semantic and geometric context under a larger effective field of view. Following the U-shaped spherical hierarchy of SphereUFormer [18], it operates on the downsampled icosphere at rank r'=r−1. Given the fused input feature $\mathbf{y}_0$ from Eq. 5, the stream first encodes features through a sequence of Adaptive Spherical Attention Down Modules (AdaSpA Down Modules) *encoders* with progressive graph *downsampling*, thereby increasing the receptive field while reducing spatial resolution. A *bottleneck* stage at the coarsest icosphere resolution aggregates global scene-level context. A series of AdaSpA Up Modules *decoders* then progressively upsamples the features and combines them with encoder features through skip connections:

$$\mathbf{y}_i^{\text{dec}} = f_i^{\text{dec}}([\text{LN}(\mathbf{y}_i^{\text{up}}), \text{LN}(\mathbf{e}_{S-1-i})]), \tag{6}$$

where $\mathbf{y}_i^{\text{dec}}$ is the decoder feature at stage i, $\mathbf{y}_i^{\text{up}}$ is the upsampled feature, and $\mathbf{e}_{S-1-i}$ is the corresponding encoder skip feature. Here, $[\cdot,\cdot]$ denotes channel-wise concatenation, LN denotes layer normalization, and f_i^{dec} is the decoder block. The resulting feature $\mathbf{y}_L \in \mathbb{R}^{N'\times d}$ is finally lifted back to the full icosphere rank r and fused with the Acuity Stream for dense prediction.

Acuity-Latent Fusion Head. The final outputs combine Lateral and Acuity predictions through learned gates:

$$\hat{\mathbf{y}} = \mathbf{h}_{\text{seg}}(\mathbf{y}_L) + \sigma(g_{\text{seg}})\,\mathbf{h}^a_{\text{seg}}(\mathbf{F}_{\text{acuity}}), \qquad \hat{\mathbf{s}} = \mathbf{h}_{\text{sdf}}(\mathbf{y}_L) + \sigma(g_{\text{sdf}})\,\mathbf{h}^a_{\text{sdf}}(\mathbf{F}_{\text{acuity}}) \tag{7}$$

where $\hat{\mathbf{y}} \in \mathbb{R}^{N\times K}$ denotes the final semantic logits used for segmentation prediction, $\hat{\mathbf{s}} \in \mathbb{R}^{N\times K}$ denotes the predicted per-class SDF used to compute $\mathcal{L}_{\text{SDF}}$, $\mathbf{y}_L$ is the final Lateral decoder feature, and $\mathbf{h}^a_{\text{seg}}$ and $\mathbf{h}^a_{\text{sdf}}$ denote the Acuity Stream prediction heads.

Method	Mild					Robust					Stress				
	Rot.	Scale	Ori.	Trans.	View.	Rot.	Scale	Ori.	Trans.	View.	Rot.	Scale	Ori.	Trans.	View.
SFSS	50.40	55.47	47.02	59.45	48.51	35.65	47.84	34.14	52.84	33.00	24.66	42.39	24.20	46.11	20.26
SGAT4PASS	59.92	62.62	56.27	65.05	57.19	37.59	55.83	36.11	56.64	32.61	18.56	49.22	18.74	47.68	12.50
Elite360D	54.21	58.92	48.23	64.17	51.51	31.37	47.02	28.08	56.00	29.26	17.01	38.12	15.05	47.77	14.52
HEAL-SWIN	36.94	40.71	33.13	43.61	32.18	25.42	34.08	24.23	34.60	20.54	17.01	30.39	15.61	28.65	12.40
SphereUFormer	64.01	69.33	59.16	67.23	55.27	36.29	60.13	33.89	57.21	30.76	18.96	51.93	17.61	48.29	14.66
SO3UFormer	65.85	65.30	60.01	66.70	59.81	46.88	55.11	41.95	53.94	36.64	27.07	46.53	24.01	41.70	19.19
AdapToPASS (Ours)	**70.93**	**71.44**	**66.12**	**72.43**	**65.90**	**51.52**	**62.23**	**46.61**	**62.90**	**45.20**	**30.43**	**54.10**	**27.05**	**53.17**	**25.84**

Table 1: **mIoU comparison under unseen spherical transformations on the Stanford2D3D benchmark.** Results are reported across mild, robust, and stress settings, including rotation (Rot.), Scale, Orientation Shifts (Ori.), Translation (Trans.), and Viewpoint Shifts (View.). *Higher is better*.

Signed Distance Field (SDF) based Boundary Supervision. Region-level cross-entropy supervision does not penalize boundary ambiguities, leading the model free to produce diffuse predictions at semantic transitions [57]. Biological perception is strongly boundary-aware, maintaining sharp delineations between regions even under viewpoint changes. We thus introduce SDF-based boundary supervision, which encodes semantic boundary on the icosphere graph through per-class signed distance targets used as auxiliary supervision during training.

Given semantic labels $\mathbf{y} \in \{0, \dots, K\}^N$, we identify boundary nodes and compute the graph distance d_i^{∂} from each node to its nearest boundary, where negative values denote nodes inside the class region and positive values denote outside. We attach lightweight SDF heads to the Lateral and Acuity features, and supervise them with a truncated regression loss:

$$\mathcal{L}_{\text{SDF}} = \frac{1}{|\mathcal{M}|} \sum_{(i,k)\in\mathcal{M}} \min\Big(|\hat{s}_i^k - s_i^{k,\text{gt}}|, \delta\Big), \tag{8}$$

where $\mathcal{L}_{\text{SDF}}$ is the truncated SDF regression loss, $\mathcal{M}$ is the set of valid non-background node-class pairs, $\hat{s}_i^k$ and $s_i^{k,\text{gt}}$ are the predicted and target signed distances, and δ is the truncation threshold.

We further use a sign-alignment loss to encourage consistency between semantic predictions and the SDF-induced inside-outside structure:

$$\mathcal{L}_{\text{align}} = \frac{1}{|\mathcal{M}|} \sum_{(i,k)\in\mathcal{M}} \Big|\text{softmax}(\hat{\mathbf{y}}_i)_k - \mathbb{1}[\hat{s}_i^k < 0]\Big|, \tag{9}$$

where $\mathcal{L}_{\text{align}}$ is the alignment loss, $\hat{\mathbf{y}}_i \in \mathbb{R}^K$ is the semantic logit vector at node i, and $\mathbb{1}[\hat{s}_i^k < 0]$ is the indicator that the predicted SDF places node i inside class k.

3.3 Training Objective and Adaptive Inference

AdapToPASS is trained with a joint objective that combines semantic segmentation, boundary-aware SDF supervision, semantic-SDF consistency, and ambiguity regularization:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{Lov}}\,\mathcal{L}_{\text{Lov}} + \lambda_{\text{SDF}}\,\mathcal{L}_{\text{SDF}}(\hat{\mathbf{S}}, \mathbf{S}^{\text{gt}}(\mathbf{D}^{\partial}, \mathbf{Y})) + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}}(\hat{\mathbf{Y}}, \hat{\mathbf{S}}) + \lambda_c\,\bar{c}. \tag{10}$$

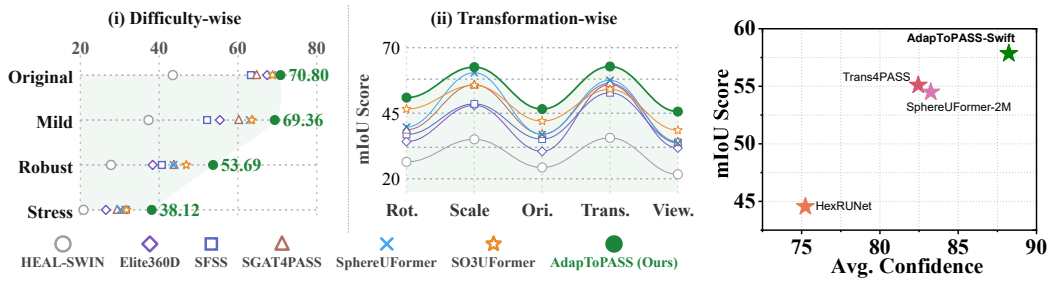
Here, $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{Lov}}$ supervise semantic prediction, $\mathcal{L}_{\text{SDF}}$ provides boundary-aware geometric supervision, $\mathcal{L}_{\text{align}}$ encourages consistency between semantic logits and predicted SDF structure, and $\bar{c}$ regularizes the learned ambiguity signal. Detailed formulations are provided in Appendix A.2.

AdapToPASS remains adaptive at during inference because the contextual ambiguity signal is computed within the forward pass rather than introduced as a training-only auxiliary variable. For each input panorama, AdaSpA predicts a token-wise ambiguity score c_i from the current spherical feature $\mathbf{x}_i$ and uses it to modulate the geodesic distance scale in Eq. (4). This yields an input-dependent attention bias $b_{ij}^h(c_i)$, allowing each token to adjust its effective aggregation range on the sphere without test-time optimization or gradient updates.

4 Experiments and Evaluations

4.1 Settings and Implementation Details

Evaluation and Experimental Settings. We evaluate AdapToPASS on two panoramic semantic segmentation benchmarks: *Stanford2D3D* [34] and *WildPASS* [35, 11] (*80:10:10* for *train:val:test*). Stanford2D3D is evaluated using all semantic classes, following the full indoor scene parsing setting.



(a) (b)

Figure 5: **(a)** Analysis on Stanford2D3D under unseen spherical transformations. (i) Difficulty-wise mIoU comparison across difficulties. (ii) Transformation-wise mIoU comparison under Rotation, Scale, Orientation Shift (Ori.), Translation, and Viewpoint Shift (View). **(b)** Comparison of AdapToPASS-Swift (our 2M param variant) with lightweight panoramic segmentation baselines.

For WildPASS, we evaluate on the three most frequent foreground classes, which together cover approximately 92% of the annotated labels. This reduced WildPASS label space is used because our primary objective is to study robustness and transformation invariance under spherical perturbations, rather than recognition of rarely occurring small objects. We provide further discussion of this choice in the Appendix B.2. All experiments were conducted on a system equipped with 4× NVIDIA GeForce RTX 5090 GPUs (32 GB VRAM each). The hyperparameter details are provided in Appendix B.4.

Unseen Spherical Transformations. To evaluate robustness beyond canonical panoramic views, we construct transformed evaluation variants of Stanford2D3D and WildPASS using diverse spherical transformations. For a fair evaluation of robustness, *all methods are trained only on the original, untransformed training sets*, without any spherical-transformation-based data augmentation. The transformed variants are used exclusively for evaluation. We consider five transformation families: *Rotation*, *Scale*, *Translation*, *Orientation Shift*, and *Viewpoint Shift*. Each family is further divided into three difficulty regimes: *mild*, *robust*, and *stress*, with increasingly larger geometric perturbations. Further details are provided in Appendix B.1.

Model	Params (M)	Stanford2D3D mAcc.↑	Stanford2D3D mIoU↑	Stanford2D3D Conf.↑	WildPASS mAcc.↑	WildPASS mIoU↑	WildPASS Conf.↑
SFSS	15.28	74.22	63.29	85.97	77.85	65.27	81.06
SGAT4PASS	15.22	73.13	64.75	91.76	72.15	56.77	82.35
Elite360D	15.41	76.57	67.42	88.10	74.36	65.40	59.65
HEAL-SWIN	14.06	72.60	43.46	88.01	64.64	48.20	81.70
SphereUFormer	14.77	78.13	68.60	90.34	79.15	63.38	74.89
SO3UFormer	14.92	78.68	68.79	91.07	72.11	57.74	72.68
AdapToPASS (Ours)	14.78	**83.08**	**70.80**	**96.42**	**81.44**	**69.58**	**87.31**

Table 2: **Comparison with state-of-the-art panoramic semantic segmentation methods on Stanford2D3D and WildPASS.** We report model parameters, mean accuracy (mAcc), mean IoU (mIoU), and mean prediction confidence (Conf.).

Baseline Methods. We compare AdapToPASS with representative ERP-based, spherical, and hybrid panoramic segmentation methods. For ERP baselines, we include SFSS as a strong RGB-only method and SGAT4PASS [19] as an ERP-based approach with spherical geometry-aware modeling. For spherical baselines, we compare with HEAL-SWIN [17], SphereUFormer [18], and SO3UFormer [20], which operate directly on spherical discretizations. We further include Elite360D [48], a hybrid method combining ERP features with a low-resolution icosahedral projection. Additional implementation details are provided in Appendix B.3.

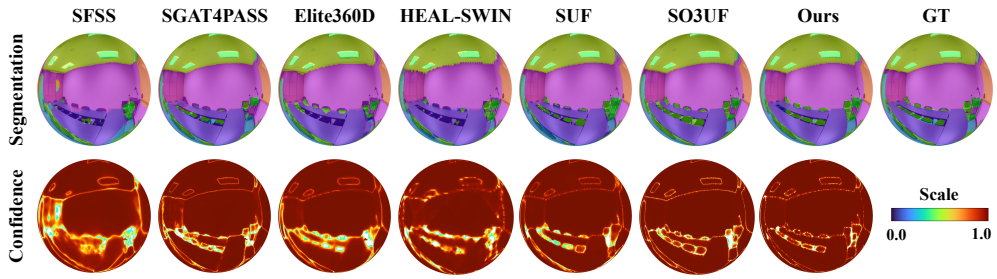


Figure 6: **Qualitative comparison on Stanford2D3D.** We compare segmentation predictions and their corresponding confidence maps across representative panoramic segmentation methods. AdapToPASS produces more accurate semantic regions with sharper boundaries while maintaining high confidence (SUF: SphereUFormer, SO3UF: SO3UFormer). *Zoom in for a better view.*

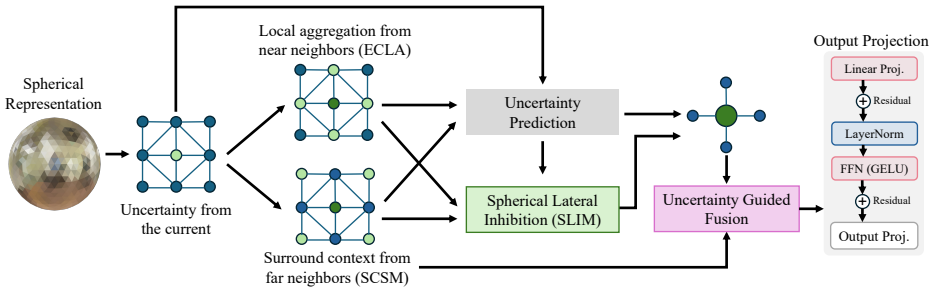


Figure 7: **Overview of the AdaSpX Module.** AdaSpX replaces the AdaSpA blocks in AdapToPASS to enable efficient spherical semantic segmentation.

4.2 Results and Discussion

We structure our evaluation around the following experimental questions (EQs):

EQ-1: Does modeling contextual and geometric ambiguity within a Spherical Transformer framework improve the accuracy and consistency of panoramic semantic segmentation?

EQ-2: Does the bio-inspired ambiguity-aware design of AdapToPASS enable more reliable generalization than canonical methods on unseen spherical transformations?

EQ-3: What is the impact of AdaSpA, Bifocal Spherical Representation, and boundary supervision on the overall performance of AdapToPASS?

Comparison with SOTA Methods. Table 2 and Fig 1(b-c) compares AdapToPASS against existing panoramic semantic segmentation methods on Stanford2D3D and WildPASS. We report the test set mean Intersection-over-Union (mIoU) and mean class accuracy (mAcc)[1].

[A-1] AdapToPASS achieves the best overall performance across both datasets, obtaining **70.80%** mIoU on Stanford2D3D, and **69.58%** mIoU on WildPASS. Notably, AdapToPASS produces the highest average prediction confidence (Conf.), indicating more reliable semantic predictions.

Fig 6 illustrates the qualitative result on Stanford2D3D. Refer to Appendix C for additional results.

Robustness Under Unseen Spherical Transformations. Tables 1 and 7 evaluate robustness under unseen spherical transformations on Stanford2D3D and WildPASS.

[A-2] AdapToPASS achieves the highest mIoU across all transformation types and severity levels. Figures 5a and 12 show that AdapToPASS achieves upto **+13.38%** and **+18.77% relative improvements in mIoU** on Stanford2D3D and WildPASS, respectively.

While the robust and stress settings introduce severe geometric distortions, orientation shifts, and viewpoint changes that degrade all methods, AdapToPASS consistently retains a clear margin over prior approaches. These results demonstrate that bio-inspired, ambiguity-aware adaptive spherical perception enables stronger generalization to unseen spherical transformations, making AdapToPASS **closer to human perceptual behavior**, even without transformation-specific augmentation.

Ablation Study. Table 3 presents an ablation analysis on the Stanford2D3D benchmark to evaluate the contribution of each component in AdapToPASS. We observe several interesting findings: **(I)** Removing the contextual ambiguity modeling significantly reduces the overall prediction confidence; **(II)** Removing the SDF supervision causes a notable drop in mIoU; **(III)** Using only contextual ambiguity with boundary supervision results a compatibility better mIoU and confidence.

Configuration	mIoU ↑	Conf. ↑
AdapToPASS	70.80	96.42
AdapToPASS w/o Context Ambiguity	64.97	88.47
AdapToPASS w/o Acuity Stream	66.50	92.20
AdapToPASS w/o Boundary Supervision	63.70	91.90

Table 3: Ablation study evaluating the contribution of key components in AdapToPASS.

[A-3] AdapToPASS achieves the strongest performance when contextual ambiguity modeling, Bifocal Spherical Representation, and boundary supervision are jointly integrated. Their combination enables adaptive context aggregation, geometry-aware multi-scale representation, and sharper boundary localization. Fig. 15 illustrates the effect of each of these component.

[1] mAcc denotes the average of per-class pixel accuracies with equal weighting across semantic categories, thereby reducing the dominance of large or frequent classes

Method	Mild					Robust					Stress				
	Rot.	Scale	Ori.	Trans.	View.	Rot.	Scale	Ori.	Trans.	View.	Rot.	Scale	Ori.	Trans.	View.
HexRUNet	37.84	39.01	35.91	38.29	34.01	28.94	34.62	27.53	33.39	24.60	19.98	31.06	18.45	28.67	15.68
SphereUFormer	37.40	53.00	27.10	35.50	19.80	30.40	27.20	23.40	50.70	19.80	26.40	38.10	30.30	13.30	19.10
Trans4PASS	57.40	56.83	54.42	53.78	54.08	40.10	50.49	38.63	45.94	34.12	25.42	44.31	25.66	38.44	18.87
AdapToPASS-Swift (Ours)	**62.20**	**61.80**	**57.80**	**60.10**	**55.20**	**47.80**	**54.00**	**43.80**	**50.50**	**38.90**	**31.80**	**47.00**	**28.60**	**41.70**	**25.30**

Table 4: **mIoU comparison of AdapToPASS-Swift with lightweight panoramic semantic segmentation models under unseen spherical transformations on the Stanford2D3D benchmark.** Results are reported across mild, robust, and stress settings, including rotation (Rot.), Scale, Orientation Shifts (Ori.), Translation (Trans.), and Viewpoint Shifts (View.). *Higher is better.*

Setting	Beam	Board	Bookcase	Ceiling	Chair	Clutter	Column	Door	Floor	Sofa	Table	Wall	Window
Original	77.50	77.15	65.99	90.46	60.22	54.43	49.07	74.55	88.59	60.35	61.24	77.88	83.04
Mild	72.81	79.89	68.46	85.94	57.31	56.20	52.25	74.22	80.66	54.62	60.49	80.21	78.70
Robust	53.71	60.57	53.96	74.51	44.02	42.52	31.72	58.05	64.47	34.36	45.78	71.12	63.26
Stress	34.02	40.82	40.07	56.26	33.62	31.15	20.02	42.18	43.24	18.50	30.55	60.91	44.31

Table 5: **Class-wise performance analysis on Stanford2D3D.** Per-class IoU (%) of AdapToPASS under progressively stronger transformations, compared with the original untransformed setting.

AdapToPASS-Swift. We further introduce AdapToPASS-Swift, a lightweight variant of AdapToPASS with fewer than 2M parameters. It follows the exact bio-inspired, ambiguity-aware spherical perception principle as AdapToPASS, but replaces AdaSpA with lightweight **Adaptive Spherical Context Aggregation Module** (**AdaSpX**), illustrated in Fig. 7 for efficient spherical semantic understanding. Under a matched compact-model setting, AdapToPASS-Swift generalizes better than similarly sized panoramic segmentation baselines, achieving stronger and more confident overall performance (Fig. 5b, Table 8) and robustness to unseen spherical transformations (Table 4). For implementation details and baseline setup please refer to Appendix D and B.3.

Failure Cases. Although AdapToPASS consistently outperforms prior methods under unseen spherical transformations, the performance still degrades noticeably under extremely severe perturbations, especially in the stress setting (Fig 5a). This degradation is most pronounced for *orientation shifts* and *viewpoint shifts*, where large changes in spherical layout can disrupt semantic consistency and introduce severe geometric ambiguity (Fig 8).

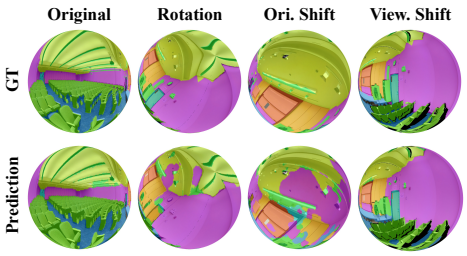


Figure 8: Failure cases of AdapToPASS on Stanford2d3d under *Stress* setting.

Table 5 presents the per-class IoU under progressively stronger geometric transformations. While performance generally decreases with transformation severity, the degradation varies across semantic categories. Structural classes such as *wall* and *ceiling* remain relatively robust, whereas *sofa* and *column* exhibit larger declines. These findings highlight the limitations of recovering discriminative semantic information under severe geometric distortions, motivating further improvements in class-specific robustness.

5 Conclusion and Future Work

In this work[2], we investigated the role of ambiguity-aware spherical perception for robust panoramic semantic segmentation. We introduced AdapToPASS, a bio-inspired Spherical Transformer framework inspired by uncertainty-driven contextual and geometric processing in biological vision. Comprehensive evaluations demonstrate that AdapToPASS consistently achieves strong segmentation performance. More importantly, it demonstrates significantly improved robustness to unseen spherical transformations compared to prior baselines.

Future Work: We aim to extend AdapToPASS in two directions: scaling ambiguity-aware spherical perception toward foundation models, including integration with large segmentation models such as SAM3, and developing lightweight variants for efficient robotic deployment. We hope this work motivates further research on ambiguity-aware spherical perception for robust robotic vision.

[2]Project Page: https://empactlab.github.io/AdapToPASS-NeurIPS-2026/

Acknowledgment

This work was supported by the MOE AcRF Tier 1 Grant under Grant No. RG160/25, MOE AcRF Tier 1 Seed Funding Grant under Grant No. RS34/24, and NTU Start-Up Grant.

References

[1] Hao Ai, Zidong Cao, Jinjing Zhu, Haotian Bai, Yucheng Chen, and Lin Wang. Deep learning for omnidirectional vision: A survey and new perspectives. *arXiv preprint arXiv:2205.10468*, 2022.

[2] Thiago LT da Silveira and Claudio R Jung. Omnidirectional visual computing: Foundations, challenges, and applications. *Computers & Graphics*, 113:89–101, 2023.

[3] Fanglue Zhang, Junhong Zhao, Yun Zhang, and Stefanie Zollmann. A survey on 360 images and videos in mixed reality: algorithms and applications. *Journal of Computer Science and Technology*, 38(3):473–491, 2023.

[4] Bui Duc Manh, Soumyaratna Debnath, Zetong Zhang, Shriram Damodaran, Arvind Kumar, Yueyi Zhang, Lu Mi, Erik Cambria, and Lin Wang. Mind meets space: Rethinking agentic spatial intelligence from a neuroscience-inspired perspective. *arXiv preprint arXiv:2509.09154*, 2025.

[5] Shriram Damodaran, Soumyaratna Debnath, Yan Wu, Wei-Yun Yau, and Addison Lin Wang. Eagor: Embodied reasoning in omni-direction. *arXiv preprint arXiv:2607.06165*, 2026.

[6] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[7] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[8] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.

[9] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021.

[10] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

[11] Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. Is context-aware cnn ready for the surroundings? panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, 30:1866–1881, 2021.

[12] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360 imagery. *Advances in neural information processing systems*, 30, 2017.

[13] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

[14] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European conference on computer vision (ECCV)*, pages 518–533, 2018.

[15] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 707–722, 2018.

[16] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Philip HS Torr, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8549–8567, 2024.

[17] Oscar Carlsson, Jan E Gerken, Hampus Linander, Heiner Spieß, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Heal-swin: A vision transformer on the sphere. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6067–6077, 2024.

[18] Yaniv Benny and Lior Wolf. Sphereuformer: A u-shaped transformer for spherical 360 perception. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 940–950, 2025.

[19] Xuewei Li, Tao Wu, Zhongang Qi, Gaoang Wang, Ying Shan, and Xi Li. Sgat4pass: Spherical geometry-aware transformer for panoramic semantic segmentation. *arXiv preprint arXiv:2306.03403*, 2023.

[20] Qinfeng Zhu, Yunxi Jiang, and Lei Fan. So3uformer: Learning intrinsic spherical features for rotation-robust panoramic segmentation. *arXiv preprint arXiv:2602.22867*, 2026.

[21] Xiangfu Yang, Weiping Yang, Yinghua Yu, Yoshimichi Ejima, and Jiajia Yang. Increasing visual uncertainty modulates multisensory decision-making. *Cortex*, 2025.

[22] Jonathan Vacher, Claire Launay, Pascal Mamassian, and Ruben Coen-Cagli. Measuring uncertainty in human visual segmentation. *PLoS Computational Biology*, 19(9):e1011483, 2023.

[23] Soumyaratna Debnath, Ashish Tiwari, Kaustubh Sadekar, and Shanmuganathan Raman. Rasp: revisiting 3d anamorphic art for shadow-guided packing of irregular objects. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5849–5858. IEEE, 2025.

[24] Soumyaratna Debnath, Bui Duc Manh, Zinan Liu, and Lin Wang. Llmind: Bio-inspired training-free adaptive visual representations for vision-language models. *arXiv preprint arXiv:2603.14882*, 2026.

[25] Zinan Liu, Eric Zheng, Soumyaratna Debnath, Hao Shi, Ling Xiao, and Lin Wang. Vl2spike: Spike-driven distillation from vlms for low-power visual perception in embodied ai. *arXiv preprint arXiv:2606.15898*, 2026.

[26] Michael F Land and Dan-Eric Nilsson. *Animal eyes*. Oxford University Press, 2012.

[27] Eric L Schwartz. Computational anatomy and functional architecture of striate cortex: a spatial mapping approach to perceptual coding. *Vision research*, 20(8):645–669, 1980.

[28] Eric Warrant. Vision in the dimmest habitats on earth. *Journal of Comparative Physiology A*, 190(10):765–789, 2004.

[29] Mujeeb Ur Rehman Parrey. From cats to the cortex: Unravelling the hierarchical processing system of vision and brain plasticity. *Cureus*, 16(9), 2024.

[30] Chetan Nandakumar. *Invariance in Human Visual Perception*. PhD thesis, UC Berkeley, 2011.

[31] Yena Han, Gemma Roig, Gad Geiger, and Tomaso Poggio. Scale and translation-invariance for novel objects in human vision. *Scientific reports*, 10(1):1411, 2020.

[32] Edmund T Rolls. Learning invariant object and spatial view representations in the brain using slow unsupervised learning. *Frontiers in Computational Neuroscience*, 15:686239, 2021.

[33] Jun-ya Okamura, Daisuke Fukano, Keisuke Murakami, and Gang Wang. View-invariant object representation in anterior and posterior inferotemporal cortex: A machine learning approach. *IBRO Neuroscience Reports*, 2025.

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

[35] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1376–1386, 2021.

[36] Soumyaratna Debnath, Ashish Tiwari, and Shanmuganathan Raman. Scribgen: generating scribble art through metaheuristics. In *SIGGRAPH Asia 2024 Art Papers*, pages 1–9. 2024.

[37] Sakshi Sakshi, Rohit Pratyush Behera, Hongyu Zhou, Yifan Wang, and Hortense Le Ferrand. Robotic materials with bioinspired microstructures for high sensitivity and fast actuation. *Advanced Science*, 13(15):e09739, 2026.

[38] Soumyaratna Debnath. *Computational Art and Pose: Solutions for 3D Packing, Scribble Art and Pose Lifting*. PhD thesis, Indian Institute of Technology Gandhinagar, 2025.

[39] Soumyaratna Debnath, Abhirup Deb, Sourav De, and Sandip Dey. Modified harris hawk optimization algorithm for multilevel image thresholding. *Hybrid Computational Intelligent Systems*, pages 291–310, 2023.

[40] Weiming Zhang, Yexin Liu, Xueye Zheng, and Lin Wang. Goodsam: Bridging domain and capacity gaps via segment anything model for distortion-aware panoramic semantic segmentation. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 28264–28273, 2024.

[41] Weiming Zhang, Dingwen Xiao, Aobotao Dai, Yexin Liu, Tianbo Pan, Shiqi Wen, Lei Chen, and Lin Wang. Leader360v: The large-scale, real-world 360 video dataset for multi-task learning in diverse environment. *ArXiv*, abs/2506.14271, 2025.

[42] Hao Ai, Zidong Cao, and Lin Wang. A survey of representation learning, optimization strategies, and applications for omnidirectional vision: H. ai et al. *International Journal of Computer Vision*, 133(8):4973–5012, 2025.

[43] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelhagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16917–16927, 2022.

[44] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: panorama transformer for indoor 360 depth estimation. In *European Conference on Computer Vision*, pages 195–211. Springer, 2022.

[45] Carlos Esteves, Christine Allen-Blanchette, Ameesh Makadia, and Kostas Daniilidis. Learning so (3) equivariant representations with spherical cnns. In *Proceedings of the european conference on computer vision (ECCV)*, pages 52–68, 2018.

[46] Taco Cohen, Maurice Weiler, Berkay Kicanaoglu, and Max Welling. Gauge equivariant convolutional networks and the icosahedral cnn. In *International conference on Machine learning*, pages 1321–1330. PMLR, 2019.

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

[48] Hao Ai and Lin Wang. Elite360d: Towards efficient 360 depth estimation via semantic-and distance-aware bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9926–9935, 2024.

[49] Sharon Zhou, Jiequan Zhang, Hang Jiang, Torbjörn Lundh, and Andrew Y Ng. Data augmentation with mobius transformations. *Machine Learning: Science and Technology*, 2(2):025016, 2021.

[50] Prakash Chandra Chhipa, Meenakshi Subhash Chippa, Kanjar De, Rajkumar Saini, Marcus Liwicki, and Mubarak Shah. Möbius transform for mitigating perspective distortions in representation learning. In *European Conference on Computer Vision*, pages 345–363. Springer, 2024.

[51] Zidong Cao, Jinjing Zhu, Weiming Zhang, Hao Ai, Haotian Bai, Hengshuang Zhao, and Lin Wang. Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 982–992, 2025.

[52] Thomas W Mitchel, Noam Aigerman, Vladimir G Kim, and Michael Kazhdan. Möbius convolutions for spherical cnns. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022.

[53] Mukai Yu, Mosam Dabhi, Liuyue Xie, Sebastian Scherer, and László A Jeni. Unified spherical frontend: Learning rotation-equivariant representations of spherical images from any camera. *arXiv preprint arXiv:2511.18174*, 2025.

[54] Duc Cao Dinh, Seok Joon Kim, and Kyusung Cho. Geometric exploitation for indoor panoramic semantic segmentation. *Advances in Neural Information Processing Systems*, 37:26355–26376, 2024.

[55] Xing Hu, Yi An, Cheng Shao, and Pan Qin. A new deep distortion convolutional neural network for semantic segmentation of panoramic images. In *Journal of Physics: Conference Series*, volume 1873, page 012006. IOP Publishing, 2021.

[56] Mengfei Duan, Yuheng Zhang, Yihong Cao, Fei Teng, Kai Luo, Jiaming Zhang, Kailun Yang, and Zhiyong Li. Panoramic out-of-distribution segmentation, 2025.

[57] Chi Wang, Yunke Zhang, Miaomiao Cui, Peiran Ren, Yin Yang, Xuansong Xie, Xian-Sheng Hua, Hujun Bao, and Weiwei Xu. Active boundary loss for semantic segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 2397–2405, 2022.

[58] Yongjie Chen, Hongmin Liu, Haoran Yin, and Bin Fan. Building vision transformers with hierarchy aware feature aggregation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5908–5918, 2023.

[59] Yang Chen, Yueqi Duan, Runzhong Zhang, and Yap-Peng Tan. Adaptive margin contrastive learning for ambiguity-aware 3d semantic segmentation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[60] Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

[61] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018.

[62] Pan Zhou, Xingyu Xie, Zhouchen Lin, and Shuicheng Yan. Towards understanding convergence and generalization of adamw. *IEEE transactions on pattern analysis and machine intelligence*, 46(9):6486–6493, 2024.

[63] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3213–3223, 2016.

[64] Yuanpeng Tu, Boshen Zhang, Yuxi Li, Liang Liu, Jian Li, Yabiao Wang, Chengjie Wang, and Cai Rong Zhao. Learning from noisy labels with decoupled meta label purifier. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19934–19943, 2023.

[65] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3533–3541, 2019.

A Additional Methodological Details

A.1 Icosphere Representation

Icosphere construction. AdapToPASS discretises the unit sphere $\mathbb{S}^2$ using a subdivided icosahedron (*icosphere*). Starting from the regular icosahedron ($V_0 = 12$ vertices, $F_0 = 20$ triangular faces), each subdivision step replaces every triangle with four smaller triangles whose new mid-edge vertices are projected onto $\mathbb{S}^2$. After r such steps, a rank-r icosphere has

$$V_r = 10 \cdot 4^r + 2 \quad \text{vertices} \qquad \text{and} \qquad F_r = 20 \cdot 4^r \quad \text{faces}, \tag{11}$$

each subdivision increasing resolution by a factor of four.

ERP-to-icosphere sampling. Input panoramas are stored as equirectangular projections (ERP), where pixel (u, v) in an image of width W and height H corresponds to azimuth $\theta = (u/W) \cdot 360 - 180$ and polar angle $\phi = (v/H) \cdot 180$. To lift an ERP image onto the icosphere, we first convert each vertex normal $\mathbf{p}_i = (x_i, y_i, z_i)$ to spherical coordinates:

$$\phi_i = \arccos(z_i) \in [0^\circ, 180^\circ], \qquad \theta_i = \operatorname{atan2}(y_i,\, x_i) \in [-180^\circ, 180^\circ]. \tag{12}$$

These are mapped to the normalized grid coordinates expected by bilinear sampling:

$$\tilde{u}_i = \frac{\theta_i}{180}, \qquad \tilde{v}_i = \frac{2\phi_i}{180} - 1, \tag{13}$$

so that $(\tilde{u}_i, \tilde{v}_i) \in [-1, 1]^2$ spans the full ERP canvas. Each vertex then receives its feature vector by sampling the ERP image at $(\tilde{u}_i, \tilde{v}_i)$ using bilinear interpolation for RGB and depth, and nearest-neighbour interpolation for discrete semantic labels, with border padding at the poles. The result is a set of V_r feature vectors $\{\mathbf{f}_i\}_{i=1}^{V_r}$ that serve as the input sequence to the AdapToPASS encoder. At rank 7 (the default), this yields 163,842 vertices, providing sufficient angular resolution to resolve fine semantic boundaries while remaining computationally tractable.



Figure 9: **Icosphere representations at different ranks.** The same panoramic scene is shown using different icosphere subdivision levels, ranging from Rank 2 (162 vertices) to Rank 7 (163,842 vertices). Higher ranks provide finer spherical resolution and preserve more visual detail.

Hierarchical downsampling. A key structural property of the icosphere is that the vertex set at rank $r-1$ is a strict subset of the vertex set at rank r: every vertex of the coarser mesh reappears verbatim after subdivision. AdapToPASS exploits this to define a parameter-free *center-select* downsampling operator. To move from rank r to rank $r-1$, we precompute, for each of the V_{r-1} coarse vertices, the index of the matching fine vertex by exact coordinate lookup, and then simply gather those entries: $\mathbf{x}^{(r-1)} = \mathbf{x}^{(r)}[\mathcal{I}_\downarrow]$, where $\mathcal{I}_\downarrow \in \mathbb{Z}^{V_{r-1}}$ is the precomputed index map. No learnable spatial parameters are introduced; the spatial information is compressed purely by discarding the $V_r - V_{r-1}$ new vertices that were added during the last subdivision. Following each spatial downsampling step, a linear projection doubles the channel dimension from C to $2C$, so that the total feature capacity is approximately preserved across scales.

Hierarchical upsampling. The decoder restores spatial resolution one rank at a time using a geometry-aware *interpolate* upsample operator. When going from rank r to rank $r+1$, the V_r vertices that already existed at the coarser rank inherit their features directly (identity copy). The $V_{r+1} - V_r$ new mid-edge vertices, which are introduced by the subdivision step and therefore have no direct

coarse counterpart, receive features by averaging the two coarse-rank edge-adjacent vertices:

$$\mathbf{x}_i^{(r+1)} = \begin{cases} \mathbf{x}_i^{(r)} & i < V_r, \\ \dfrac{\mathbf{x}_{\ell(i)}^{(r)} + \mathbf{x}_{m(i)}^{(r)}}{2} & i \geq V_r, \end{cases} \tag{14}$$

where $\ell(i)$ and $m(i)$ are the indices of the two rank-r neighbors of new vertex i, determined once at construction time from the icosphere adjacency. This respects the subdivision geometry: new vertices lie at the midpoints of coarse edges, so averaging the two endpoints is the exact linear interpolation on the sphere. Preceding each spatial upsampling step, a linear projection halves the channel dimension from $2C$ to C.

A.2 Training Objective and Optimization

AdapToPASS is trained end-to-end with a unified objective combining semantic segmentation supervision, SDF-based boundary terms, and an ambiguity regularizer:

$$\mathcal{L}_{\text{overall}} = \mathcal{L}_{\text{CE}} + \lambda_{\text{Lov}}\,\mathcal{L}_{\text{Lov}} + \lambda_{\text{SDF}}\,\mathcal{L}_{\text{SDF}}\Big(\hat{\mathbf{S}},\, \mathbf{S}^{\text{gt}}(\mathbf{D}^{\partial}, \mathbf{Y})\Big) + \lambda_{\text{align}}\,\mathcal{L}_{\text{align}}\Big(\hat{\mathbf{Y}}, \hat{\mathbf{S}}\Big) + \lambda_c\,\bar{c}. \tag{15}$$

Here $\mathbf{Y}$ denotes the ground-truth semantic labels, $\hat{\mathbf{Y}}$ the predicted segmentation logits, $\mathbf{D}^{\partial}$ the unsigned geodesic boundary-distance field on the icosphere graph, $\mathbf{S}^{\text{gt}}(\mathbf{D}^{\partial}, \mathbf{Y})$ the per-class signed distance field derived from $\mathbf{D}^{\partial}$ and $\mathbf{Y}$, and $\hat{\mathbf{S}}$ the corresponding SDF predicted by a dedicated head. The SDF losses $\mathcal{L}_{\text{SDF}}$ and $\mathcal{L}_{\text{align}}$ are defined in eqs. (8) and (9) with $\lambda_{\text{SDF}} = 0.5$ and $\lambda_{\text{align}} = 0.1$.

The segmentation terms $\mathcal{L}_{\text{CE}}$ and $\mathcal{L}_{\text{Lov}}$ supervise the semantic head directly. $\mathcal{L}_{\text{CE}}$ is a class-balanced cross-entropy loss with effective-number-of-samples class weights [60], excluding the background class. $\mathcal{L}_{\text{Lov}}$ is the Lovász-Softmax loss [61], a tight differentiable surrogate for the mean intersection-over-union restricted to non-background classes, weighted by $\lambda_{\text{Lov}} = 1.0$.

The final term $\lambda_c\,\bar{c}$ regularises the Contextual-Ambiguity-Aware Geodesic Bias (CxAGB) modules. At each attention layer, a small MLP predicts a per-node ambiguity scalar $u_i = \sigma(\text{MLP}(\text{LN}(\mathbf{x}_i))) \in [0, 1]$ that continuously rescales the geodesic attention window: $s_i = s_{\min} + (s_{\max} - s_{\min})\,u_i$, with $s_{\min} = 0.8$ and $s_{\max} = 1.6$. $\bar{c}$ is the mean of u_i across all nodes and all CxAGB layers.

Optimization. AdapToPASS is trained with AdamW [62] using an initial learning rate of 10^{-4}, a minimum learning rate of 10^{-6}, weight decay of 0.01, and a warmup-cosine schedule with 10 warmup epochs over 400 total epochs. Linear weights are initialised with a truncated normal distribution ($\sigma = 0.02$); layer-normalisation parameters are initialised with weight 1 and bias 0. **No transformation-specific data augmentation is applied to AdapToPASS during training**.

B Additional Experiment Details

B.1 Unseen Spherical Transformations

We generate unseen transformed evaluation sets for Stanford2D3D and WildPASS to test robustness under spherical geometric perturbations. The transformation families are denoted as *Rotation*, *Scale*, *Translation*, *Orientation Shift*, and *Viewpoint Shift*. *Rotation* applies spherical yaw, pitch, and roll perturbations. *Scale* is produced by applying a spherical Möbius transformation, which conformally warps the sphere to simulate zoom while preserving local angular structure [49]. We use a Möbius zoom factor s, where $s > 1$ concentrates content around the viewing direction to create zoom-in, and $s < 1$ spreads content over a wider spherical region to create zoom-out. *Translation* applies depth-based camera shifts along the 3D axes. *Orientation Shift* combines spherical rotation with Möbius zoom, while *Viewpoint Shift* combines spherical rotation with depth-based translation.

For Stanford2D3D, we evaluate all five transformation families. For WildPASS, we evaluate *Rotation*, *Scale*, and *Orientation Shift*; *Translation* and *Viewpoint Shift* are not used because depth-based viewpoint translation requires paired depth panoramas. For each transformation family, we define three difficulty regimes: *mild*, *robust*, and *stress*. The *mild* regime represents small perturbations close to nominal viewing conditions, *robust* represents moderate viewpoint and scale changes, and *stress* represents stronger geometric shifts for out-of-distribution robustness testing. The parameter candidates used for each transformation family and difficulty regime are summarized in Table 6.

Table 6: **Difficulty levels for unseen spherical transformations.** Each level defines progressively stronger perturbations for rotation, scale, translation, orientation shift, and viewpoint shift.

Type	Parameter	Mild	Robust	Stress
Rotation	Yaw	$0°, 45°, 90°, 135°, 180°$	$0°{:}30°{:}330°$	$0°{:}30°{:}330°$
	Pitch	$-10°, -5°, 0°, 5°, 10°$	$-30°, -20°, -10°, 0°, 10°, 20°, 30°$	$-60°, -45°, -30°, 0°, 30°, 45°, 60°$
	Roll	$-10°, 0°, 10°$	$-30°, -15°, 0°, 15°, 30°$	$-45°, 0°, 45°$
Scale	Zoom	$0.9, 1.0, 1.1, 1.25, 1.5$	$0.8, 1.0, 1.2, 1.5, 2.0$	$0.6, 0.8, 1.0, 1.5, 2.0, 3.0$
Translation	t_x/t_y	$-0.08, -0.05, 0, 0.05, 0.08$	$-0.18, -0.12, -0.08, 0, 0.08, 0.12, 0.18$	$-0.30, -0.22, -0.16, -0.10, 0, 0.10, 0.16, 0.22, 0.30$
	t_z	$-0.04, 0, 0.04$	$-0.08, -0.05, 0, 0.05, 0.08$	$-0.12, -0.08, 0, 0.08, 0.12$
Orientation Shift	Composition	Rotation + Scale		
Viewpoint Shift	Composition	Rotation + Translation		

Generation protocol. For each source panorama, we generate transformed evaluation samples independently for each difficulty regime and transformation family. The transformation parameters are sampled from predefined severity-specific candidate sets. For *Rotation*, yaw, pitch, and roll are sampled from the corresponding angular ranges, while the scale factor is fixed to 1.0. For *Scale*, yaw, pitch, and roll are fixed to zero, and the Möbius zoom factor is sampled from the corresponding zoom set. For *Orientation Shift*, both spherical rotation and Möbius zoom are applied jointly. Identity transformations are excluded: pure rotation and rotation-composed families do not allow zero yaw, pitch, and roll simultaneously, and pure scale and rotation-scale families do not use zoom factor 1.0.



Figure 10: Examples of unseen spherical transformations.

For translation-based transformations, we use the paired depth panorama to perform a geometric viewpoint shift rather than a purely image-space warp. Each valid ERP pixel is first lifted to a 3D point using its depth value and spherical ray direction. The camera is then translated by (t_x, t_y, t_z), and for *Viewpoint Shift*, the translated points are additionally rotated using the sampled yaw, pitch, and roll. The transformed 3D points are reprojected back to ERP coordinates, with collisions resolved using a

nearest-depth rule. Pixels that receive no valid projection are left empty and recorded in an additional valid mask. Figure 10 provides representative examples of the unseen spherical transformations used in our robustness evaluation.

B.2 WildPASS Dataset and Label Space

The full WildPASS annotation set covers 8 Cityscapes categories [63]; however, the label distribution is highly skewed. Three foreground classes, *road*, *sidewalk*, and *car*, dominate the annotated pixel budget and together account for approximately 92% of all labelled pixels across the dataset. The remaining categories each constitute a small minority and are frequently absent from individual panoramas entirely.

We therefore collapse all categories except road, sidewalk, and car into a single *background* class (label 0), which is excluded from loss computation and evaluation. This gives a 4-class label space $\{0, 1, 2, 3\}$ = {background, road, sidewalk, car} used consistently across all experiments. As shown in Fig. 11, the WildPASS annotations are highly imbalanced, with Road, Sidewalk, and Car accounting for the majority of foreground pixels.

Our primary goal on WildPASS is to evaluate robustness and transformation invariance under spherical perturbations, not general scene parsing. Restricting evaluation to the three dominant classes serves two purposes. First, it ensures that metrics are not dominated by the large number of background pixels or by classes with too few samples to yield reliable mIoU estimates. Second, road, sidewalk, and car exhibit qualitatively different geometric footprints on the sphere, a broad ground-level stripe, a narrow transitional band, and compact mid-field objects, respectively, making them collectively informative for assessing how well a model preserves spatially structured predictions under viewpoint variation. Evaluating all methods under this identical, fixed label space ensures a fair and reproducible comparison.

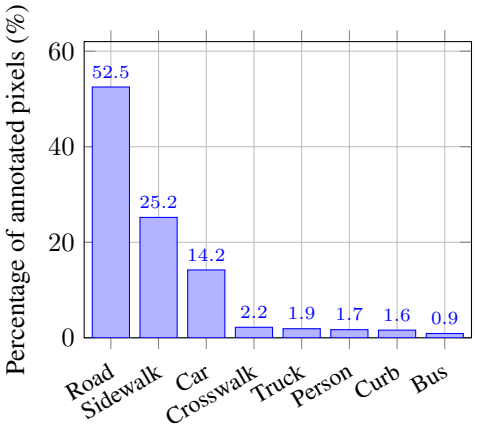


Figure 11: Class distribution for WildPASS dataset.

B.3 Baseline Implementation Details

We summarize the implementation details used for the baseline comparisons. Patch-based foundation-model approaches are excluded from the comparison due to their substantially larger parameter counts, computational cost, and different inference paradigm.

Note. For a fair comparison, all models are trained from scratch with RGB inputs only, using comparable model settings when variants exist. Predictions from different representations are evaluated on a common uniformly sampled sphere to avoid ERP-induced bias. All models are trained bsaed on their suggested epoch

Whenever the original implementation supports multiple model sizes, we choose or configure the variant whose depth, number of downsampling stages, and parameter count are closest to our model. Specifically, we use four downsampling stages and a depth of two blocks per stage whenever supported by the architecture. For grid-based models, we evaluate input resolutions of 256×512 and 512×1024; for spherical models, we choose the closest icosahedral or spherical-grid resolution in terms of sample count. All predictions are projected uniformly onto the sphere before evaluation, avoiding the latitude bias introduced by direct ERP evaluation.

SFSS [47]. SFSS is evaluated in its unimodal RGB setting so that no depth or normal modality is used. Although the original RGB-X dataloader retains an auxiliary tensor, the unimodal model uses a single-stream RGB backbone and ignores the auxiliary input. We use a SegFormer-style [10]

encoder with a DMLP decoder [64], and enable deformable convolution only in the first stage. The model is trained from scratch using AdamW with learning rate 6×10^{-5}, weight decay 0.01, polynomial learning-rate decay with power 0.9, and 10 warm-up epochs. Input panoramas are resized to 512×1024. Training uses ImageNet normalization, random horizontal flipping, multi-scale augmentation, and random 512×512 crops. Evaluation uses single-scale sliding-window inference without flip augmentation.

SGAT4PASS [19]. SGAT4PASS is reproduced using its Trans4PASS-style encoder-decoder architecture under the same RGB-only setting. The encoder follows a MiT-style hierarchy with deformable convolution enabled only in the first stage and dense decoder enhancement enabled across all stages [10]. All weights are randomly initialized. Training uses cross-entropy loss with AdamW and learning rate 8×10^{-5}. The model is trained with square crops and reprojection augmentation, including rotations along the x, y, and z axes. Evaluation follows the spherical evaluation protocol described above after projecting predictions to the sphere.

HEAL-SWIN [17]. HEAL-SWIN is evaluated as a spherical-grid Transformer based on the HEALPix representation. The model uses a Swin-style hierarchy with patch-based tokenization, shifted-window attention, and flat relative-position bias. We use the closest model scale to the other baselines, with four stages and hierarchical attention blocks. The model is trained from scratch with mixed precision using learning rate 10^{-4} and weight decay 10^{-4}. Since HEAL-SWIN operates natively on a spherical grid, its output is projected to the common spherical evaluation grid for fair comparison.

Elite360D [48]. Elite360D is evaluated using its ERP-plus-icosahedral design. The model combines an ERP branch with an icosahedral branch using a ResNet-18 backbone and face-level icosahedral features. The spherical branch uses an icosphere level chosen to match the pixel/sample count of the grid-based baselines as closely as possible. The model is trained from scratch using Adam with learning rate 10^{-4} and AMP enabled. Color augmentation, horizontal flipping, and yaw augmentation are enabled during training (matching the original setup). Final predictions are projected to the sphere for uniform evaluation.

SphereUFormer [18]. SphereUFormer is evaluated as a vertex-icosphere Transformer operating directly on the spherical domain. We use a rank-7 icosphere with four hierarchical scales, depth two per scale, encoder heads $[2, 4, 8, 16]$, and decoder heads $[16, 16, 8, 4]$. The model uses center downsampling, interpolation-based upsampling, and both absolute and relative positional encodings. Training is performed from scratch using Adam with learning rate 10^{-4}. Color augmentation is disabled, while horizontal flipping and yaw rotation remain enabled (matching the original setup). Since the model already predicts on the icosphere, its output is evaluated on the common spherical grid.

SO3UFormer [20]. SO3UFormer follows the same hierarchical icosphere structure as SphereUFormer but replaces standard spherical attention and resampling with SO(3)-aware components. It uses quadrature attention, gauge-aware relative-position bias, pool-invariant gauge mode, multiple gauge frames, area-average downsampling, geodesic-kernel upsampling, and an equivariance regularization term with weight 0.05. The model is trained from scratch in `bf16` precision using Adam with learning rate 10^{-4}. Color augmentation is disabled, while horizontal flipping and yaw rotation remain enabled (matching the original setup). Predictions are evaluated on the same uniformly sampled spherical surface as the other methods.

HexRUNet [65]. HexRUNet is implemented as a lightweight spherical encoder-decoder operating on an icosahedral level-4 representation. The model follows a compact residual U-Net design with two downsampling blocks, two decoder blocks, and a final per-face segmentation classifier. Starting from a base width of 93 channels, the encoder progressively expands the feature dimension to 186 and 372 channels before decoding back to the semantic output space. We train HexRUNet from scratch with AdamW, using learning rate 10^{-4}, weight decay 10^{-4}, and a StepLR schedule with decay factor 0.5. Training is performed with mixed precision and batch size 64. For this baseline, yaw-rotation and horizontal-flip augmentations are disabled.

Trans4PASS [43]. Trans4PASS is implemented using the compact micro configuration for panoramic semantic segmentation. The model follows a lightweight encoder-decoder design with embedding dimension 60 and backbone channel widths $[32, 64, 128, 192]$. Deformable convolution is used only in the first stage, while dense decoder enhancement is applied at all stages. We train Trans4PASS from scratch without external pretrained initialization. Optimization uses AdamW with learning rate 5×10^{-5} and cross-entropy loss. The model is trained using 1080×1080 crops and evaluated using high-resolution 2048×1024 crops following the panoramic inference protocol.

SphereUFormer-2M [18]. SphereUFormer-2M is implemented as a compact spherical Transformer operating directly on a rank-7 vertex icosphere. The model follows a three-scale hierarchical encoder-decoder design with embedding dimension 27, scale depth 2, encoder heads $[1, 3, 6]$, and decoder heads $[6, 3, 1]$. The network uses center downsampling, interpolation-based upsampling, and local spherical attention with checkpointing. We train SphereUFormer-2M from scratch using Adam with learning rate 10^{-4}, batch size 32, and drop-path rate 0.05. Following the full SphereUFormer setting, color augmentation is disabled, while horizontal flipping and yaw rotation remain enabled.

B.4 Hyperparameter Details for AdapToPASS

AdapToPASS operates on rank-7 icosphere vertices with RGB inputs and an ERP resolution of 512×256. The base embedding dimension is 32, with four encoder-decoder scales, depth 2 per encoder/decoder scale, bottleneck depth 2, encoder heads $[2, 4, 8, 16]$, bottleneck heads 16, and decoder heads $[16, 16, 8, 4]$. The Acuity Stream uses depth 1, 2 heads, and a dimension ratio of 2.

We train AdapToPASS using AdamW for 300 epochs with batch size 4 per GPU, learning rate 10^{-4}, minimum learning rate 10^{-6}, weight decay 0.01, and a warmup-cosine schedule with 10 warmup epochs. The drop-path rate is 0.05, and no explicit data augmentation is used. RGB inputs are normalized with mean 0.5 and standard deviation 0.225.

C Additional Results and Discussions

1. Table 7 reports the mIoU comparison on WildPASS under unseen spherical transformations across mild, robust, and stress settings. AdapToPASS consistently outperforms all baselines across rotation, scale, and orientation shifts, with the gains becoming especially clear under the more challenging robust and stress settings.
2. Fig. 12 further summarizes the WildPASS robustness results by averaging performance across difficulty levels and transformation families. AdapToPASS consistently achieves the highest mIoU across both difficulty levels and transformation types, showing stronger robustness under challenging spherical perturbations.
3. Fig. 13 and Fig. 14 provides a qualitative comparison on Stanford2D3D and WildPASS datasets respectively under unseen spherical transformations. While existing methods often show fragmented predictions and degraded confidence after geometric perturbations, AdapToPASS maintains more consistent semantic structure across transformations. This suggests that ambiguity-aware spherical modeling improves both prediction stability and confidence under challenging panoramic viewpoint changes.
4. **Social Impact.** AdapToPASS can support safer and more reliable panoramic perception for embodied systems such as mobile robots, assistive agents, and immersive AR/VR platforms. By improving robustness under camera motion and unseen spherical transformations, it may reduce perception failures in real-world 360° environments where stable viewpoint assumptions do not hold. Its lightweight variant further suggests potential for deployment on resource-constrained robotic systems. However, as with any scene understanding model, incorrect segmentation may still lead to downstream errors in safety-critical applications; therefore, deployment should include uncertainty monitoring, human oversight, and careful validation across diverse environments.

Method	Mild			Robust			Stress		
	Rot.	Scale	Ori.	Rot.	Scale	Ori.	Rot.	Scale	Ori.
SFSS	56.53	58.43	53.46	49.21	56.58	47.59	46.33	53.78	46.85
SGAT4PASS	64.32	66.60	62.91	58.80	65.88	56.67	51.53	63.88	52.87
Elite360D	53.14	55.09	50.00	41.26	51.17	39.55	33.02	47.44	34.23
HEAL-SWIN	50.56	52.52	48.23	46.33	50.04	44.41	43.65	48.27	42.63
SphereUFormer	61.75	63.88	61.28	55.04	62.15	53.46	42.72	58.98	41.22
SO3UFormer	62.66	62.17	61.36	61.98	61.07	58.53	61.82	60.23	56.28
AdapToPASS (Ours)	**74.09**	**75.05**	**71.52**	**74.29**	**73.09**	**69.98**	**73.74**	**71.15**	**65.70**

Table 7: **mIoU comparison under unseen spherical transformations on the WildPASS benchmark.** Results are reported across mild, robust, and stress settings, including rotation (Rot.), Scale, and Orientation Shifts (Ori.). Higher is better.

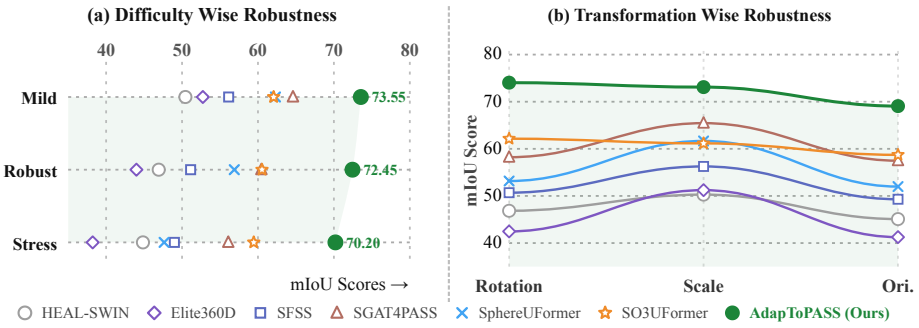


Figure 12: **Robustness analysis on WildPASS under unseen spherical transformations. (a)** Difficulty-wise average mIoU comparison across mild, robust, and stress settings. **(b)** Transformation-wise average mIoU comparison across rotation, scale, and orientation shift (Ori.).

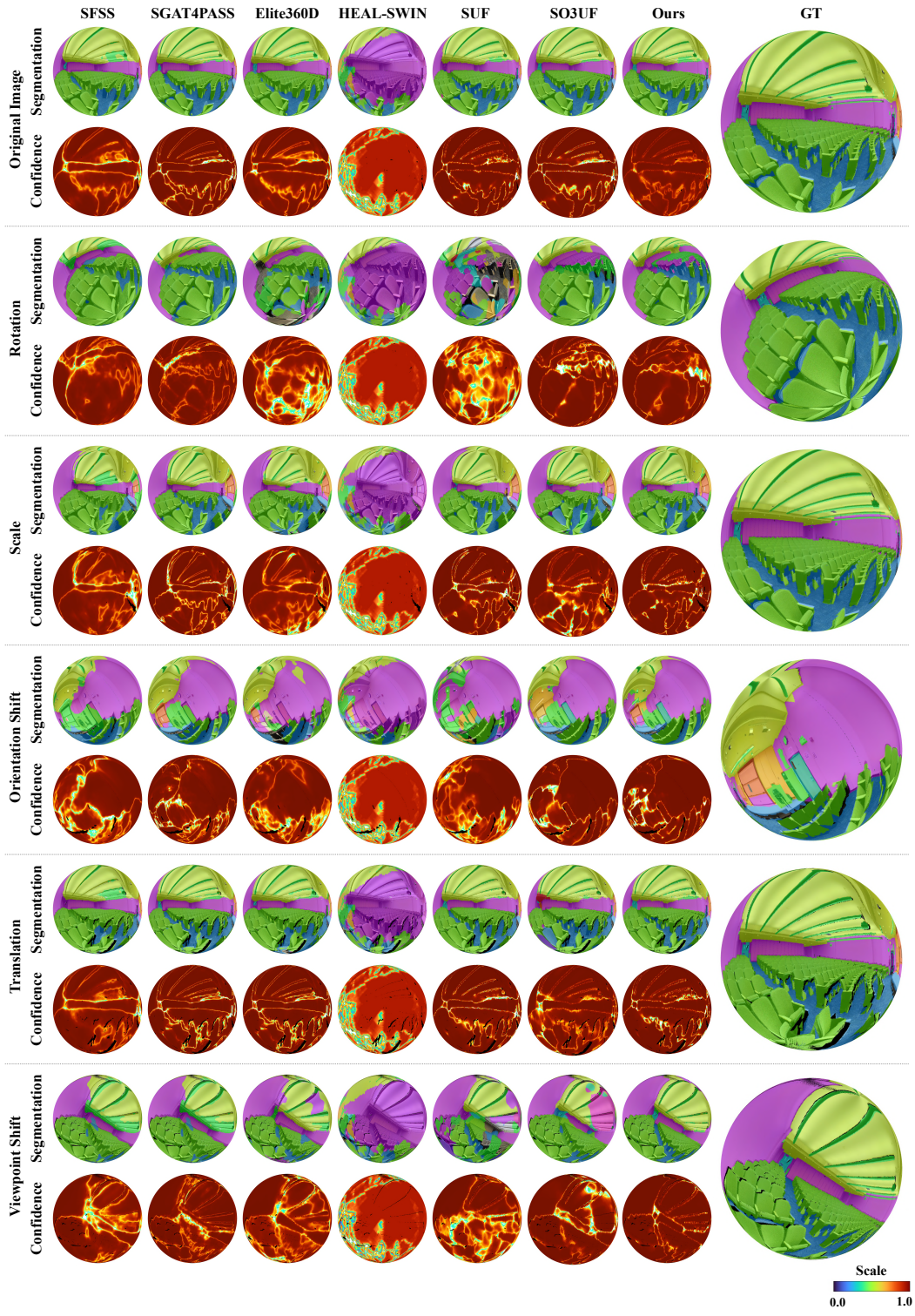


Figure 13: **Qualitative comparison on Stanford2D3D under unseen spherical transformations.** We compare segmentation predictions and confidence maps for different methods across the original image and transformed variants, including rotation, scale, orientation shift, translation, and viewpoint shift under Robust setting. Each row pair shows the predicted semantic map and the corresponding confidence map for the same input. *Zoom in for a better view.* (SUF:SphereUFormer, SO3F: SO3UFormer)

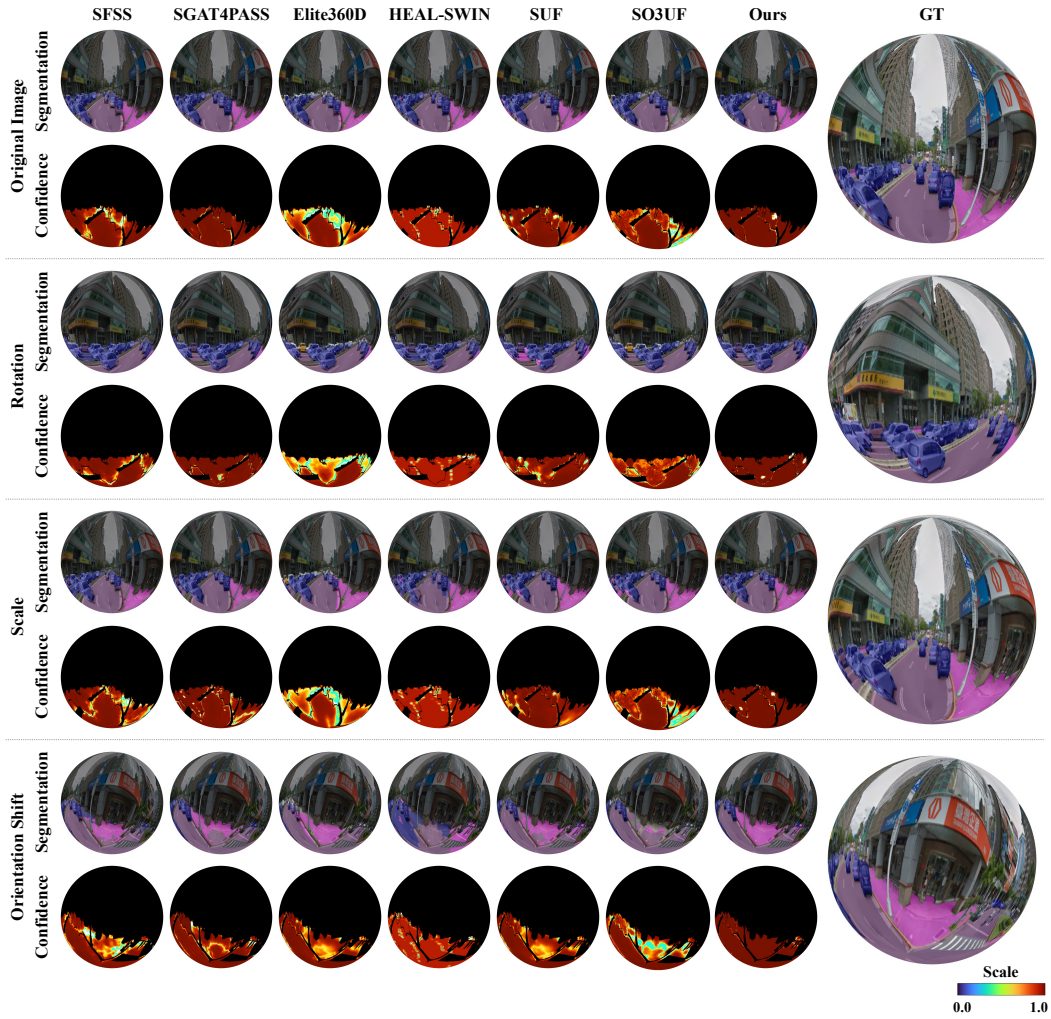


Figure 14: **Qualitative comparison on WildPASS under unseen spherical transformations.** We compare segmentation predictions and confidence maps for different methods across the original image and transformed variants, including rotation, scale, and orientation shift under Robust setting. Each row pair shows the predicted semantic map and the corresponding confidence map for the same input. *Zoom in for a better view.* (SUF:SphereUFormer, SO3F: SO3UFormer)

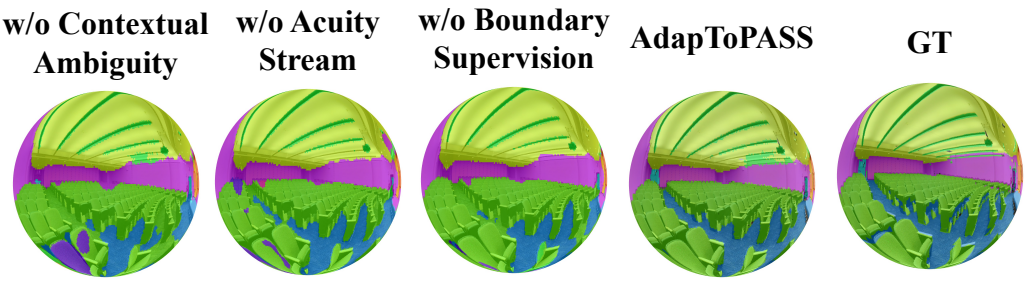


Figure 15: **Qualitative ablation study of AdapToPASS.** Removing contextual ambiguity modeling, the acuity stream, or boundary supervision degrades semantic consistency and boundary quality. The full AdapToPASS model produces more coherent predictions that better align with the ground truth.

D AdapToPASS-Swift

Overview. To further examine the efficiency and robustness tradeoff in panoramic semantic segmentation, we introduce **AdapToPASS-Swift**, a lightweight variant of AdapToPASS with **less than 2M params**, designed for resource constrained spherical perception. The goal of AdapToPASS-Swift is to preserve the key principles of ambiguity aware spherical modeling, including geometry aware local aggregation, adaptive contextual integration, and multi scale spherical perception, while reducing architectural complexity and computational overhead.

Model	Params (M)	Stanford2D3D Acc.↑	mIoU↑	Conf.↑
HexRUNet	1.99	59.79	44.55	75.24
SphereUFormer	2.04	69.42	55.10	82.45
Trans4PASS	1.99	70.60	54.50	83.25
AdapToPASS-Swift	1.99	**73.20**	**57.85**	**88.23**

Table 8: **Comparison of lightweight panoramic semantic segmentation models on Stanford2D3D.** We report average class accuracy (Acc.), mean IoU (mIoU), and mean prediction confidence (Conf.)

AdapToPASS-Swift follows the same high level design philosophy as AdapToPASS. It operates directly on the icosphere representation and maintains two complementary streams: an *Acuity stream* for preserving fine local structure and a *Lateral stream* for capturing broader contextual cues. Instead of relying on attention based token aggregation, AdapToPASS-Swift uses lightweight geometry conditioned operations to aggregate local and surround information on the sphere.

D.1 AdaSpX: Adaptive Spherical Context Aggregation

At the core of AdapToPASS-Swift is the **Adaptive Spherical Context Aggregation** (**AdaSpX**) module (illustrated in Fig. 7), an attention free spherical aggregation block for efficient ambiguity aware perception on the icosphere. Given node features $\mathbf{x} \in \mathbb{R}^{B\times N\times C}$ on the icosphere graph, AdaSpX aggregates information from two precomputed geodesic neighborhoods: a near neighborhood $\mathcal{N}_{\text{near}}(i)$ for local geometric evidence and a far neighborhood $\mathcal{N}_{\text{far}}(i)$ for wider spherical context.

Edge Conditioned Local Aggregation (ECLA). For each node v_i and neighbor $v_j \in \mathcal{N}_{\text{near}}(i)$, we construct an edge descriptor

$$\mathbf{e}_{ij} = [\mathbf{x}_i; \mathbf{x}_j; \mathbf{x}_j - \mathbf{x}_i; d_{ij}], \tag{16}$$

where d_{ij} is the geodesic distance between the corresponding icosphere vertices. A lightweight MLP predicts a channel wise gate

$$\mathbf{g}_{ij} = \sigma(\text{MLP}(\mathbf{e}_{ij})), \tag{17}$$

which modulates the neighbor message:

$$\mathbf{m}_{ij} = \mathbf{g}_{ij} \odot \frac{1}{d_{ij}+\epsilon}\mathbf{W}_v\mathbf{x}_j. \tag{18}$$

The local feature $\mathbf{l}_i$ is obtained by gated aggregation over the near neighborhood, together with a residual center projection. This provides a lightweight geometry aware mechanism for local spherical reasoning.

Surround Context Stabilization Module (SCSM). To incorporate broader contextual evidence, AdaSpX also aggregates from the far neighborhood $\mathcal{N}_{\text{far}}(i)$. A scalar edge gate is predicted for each far neighbor and used to compute a distance weighted surround representation:

$$\mathbf{s}_i = \text{SCSM}(\mathbf{x}_i, \{\mathbf{x}_j, d_{ij}\}_{j\in\mathcal{N}_{\text{far}}(i)}). \tag{19}$$

This term provides each node with a stable contextual reference, allowing ambiguous local regions to be interpreted with support from a wider spherical field of view.

Spherical Lateral Inhibition Module (SLIM). AdaSpX further applies a lightweight inhibition mechanism to suppress unstable local responses. Given the local feature $\mathbf{l}_i$, we compute a neighborhood consensus

$$\bar{\mathbf{l}}_i = \frac{1}{|\mathcal{N}_{\text{near}}(i)|}\sum_{j\in\mathcal{N}_{\text{near}}(i)} \mathbf{l}_j. \tag{20}$$

The inhibition strength is predicted as

$$\boldsymbol{\alpha}_i = \sigma\left(\mathrm{MLP}([\mathbf{l}_i; \bar{\mathbf{l}}_i; \mathbf{l}_i - \bar{\mathbf{l}}_i; u_i])\right), \tag{21}$$

where u_i denotes the uncertainty estimate at node i. The inhibited local feature is

$$\tilde{\mathbf{l}}_i = \mathbf{l}_i - \boldsymbol{\alpha}_i \odot \tanh(\mathbf{l}_i - \bar{\mathbf{l}}_i). \tag{22}$$

This operation encourages local consistency while retaining sharp transitions when the local evidence is reliable.

Uncertainty Guided Fusion. AdaSpX predicts an uncertainty embedding from the current, local, and surround features:

$$\mathbf{u}_i^{\mathrm{emb}} = \sigma\left(\mathrm{MLP}([\mathbf{x}_i; \mathbf{l}_i; \mathbf{s}_i])\right), \qquad u_i = \mathrm{mean}(\mathbf{u}_i^{\mathrm{emb}}). \tag{23}$$

The final fused representation is obtained by adaptively combining the inhibited local feature and surround context:

$$\mathbf{q}_i = \sigma\left(\mathrm{MLP}([\tilde{\mathbf{l}}_i; \mathbf{s}_i; \tilde{\mathbf{l}}_i - \mathbf{s}_i; \tilde{\mathbf{l}}_i \odot \mathbf{s}_i; \mathbf{u}_i^{\mathrm{emb}}])\right) \odot (1 - \mathbf{u}_i^{\mathrm{ch}}), \tag{24}$$

$$\mathbf{z}_i = \mathbf{q}_i \odot \tilde{\mathbf{l}}_i + (1 - \mathbf{q}_i) \odot \mathbf{s}_i. \tag{25}$$

When the predicted uncertainty is high, the fusion shifts toward broader surround context; when uncertainty is low, the block preserves stronger local evidence. The AdaSpX block is then wrapped in a standard residual FFN structure:

$$\mathbf{x}' = \mathbf{x} + \mathrm{DropPath}(\mathbf{W}_o \mathbf{z}), \qquad \mathbf{x}'' = \mathbf{x}' + \mathrm{DropPath}(\mathrm{FFN}(\mathrm{LN}(\mathbf{x}'))). \tag{26}$$

D.2 Acuity Lateral Lightweight Architecture

AdapToPASS-Swift retains the bifocal design of AdapToPASS through an *Acuity stream* and a *Lateral stream*. The Acuity stream emphasizes high resolution local structure, while the Lateral stream operates over coarser spherical resolutions to capture broad contextual information. Both streams are implemented using AdaSpX blocks, enabling efficient spherical aggregation without attention.

Acuity Stream. The Acuity stream processes features at the finest available icosphere resolution and uses compact neighborhoods to preserve local geometry. In our implementation, the Acuity stream uses smaller near and far neighborhoods:

$$K_{\mathrm{near}}^{\mathrm{acuity}} = [6, 6, 6], \qquad K_{\mathrm{far}}^{\mathrm{acuity}} = [12, 8, 6]. \tag{27}$$

This stream is responsible for maintaining fine grained spatial detail and stable local predictions.

Lateral Stream. The Lateral stream starts from a coarser spherical resolution and uses larger neighborhoods to provide wide field contextual support:

$$K_{\mathrm{near}}^{\mathrm{lateral}} = [12, 10, 8], \qquad K_{\mathrm{far}}^{\mathrm{lateral}} = [24, 18, 12]. \tag{28}$$

This stream captures semantic context over a broader field of view and supports the Acuity stream in ambiguous regions.

Uncertainty Gated Cross Stream Exchange. At each hierarchical level, the Acuity and Lateral streams exchange information through uncertainty gated bidirectional fusion. Let $\mathbf{x}_a^l$ and $\mathbf{x}_b^l$ denote the Acuity and Lateral features at level l. The module first computes cross stream messages:

$$\mathbf{m}_{b\to a}^l = \mathbf{W}_{b\to a}\mathrm{LN}(\mathbf{x}_b^l), \qquad \mathbf{m}_{a\to b}^l = \mathbf{W}_{a\to b}\mathrm{LN}(\mathbf{x}_a^l). \tag{29}$$

A fusion uncertainty signal is predicted from both streams:

$$u_{\mathrm{fus}}^l = \sigma\left(\mathrm{MLP}([\mathrm{LN}(\mathbf{x}_a^l); \mathrm{LN}(\mathbf{x}_b^l)])\right). \tag{30}$$

The gated updates are then given by

$$\mathbf{x}_a^{l,+} = \mathbf{x}_a^l + \mathbf{g}_{b\to a}^l \odot \mathbf{m}_{b\to a}^l, \tag{31}$$

$$\mathbf{x}_b^{l,+} = \mathbf{x}_b^l + \mathbf{g}_{a\to b}^l \odot \mathbf{m}_{a\to b}^l. \tag{32}$$

This fusion allows fine resolution features and wide field contextual features to interact throughout the network. Under higher uncertainty, the model can rely more strongly on the Lateral stream for contextual stabilization, while confident regions preserve stronger Acuity stream detail.

Prediction Head. The final segmentation logits are predicted from the fused Acuity features:

$$\hat{\mathbf{Y}} = h_{\text{seg}}(\mathbf{x}_a^{\text{out}}), \tag{33}$$

where h_{seg} is a lightweight segmentation head.

D.3 Training Objective

AdapToPASS-Swift is trained using the semantic segmentation objective

$$\mathcal{L}_{\text{Lite}} = w_{\text{CE}}\mathcal{L}_{\text{CE}} + w_{\text{Dice}}\mathcal{L}_{\text{Dice}}, \tag{34}$$

where the Dice term is optional and is set to zero by default. This objective keeps the training pipeline simple while allowing the model to learn ambiguity aware spherical representations through the AdaSpX forward computation.

Takeaway. AdapToPASS-Swift provides an efficient member of the AdapToPASS family for lightweight panoramic semantic segmentation. By combining adaptive spherical context aggregation, geodesic edge conditioned local reasoning, surround context stabilization, lateral inhibition, and uncertainty gated Acuity Lateral fusion, it preserves ambiguity aware spherical perception while reducing the complexity of the full model.

Hyperparameter Details for AdapToPASS-Swift. AdapToPASS-Swift operates on rank-5 icosphere vertices with RGB inputs and an ERP grid width of 1024. It uses embedding dimension 24, dimension multipliers $[1, 2, 4]$, encoder depths $[2, 2, 2]$, bottleneck depth 2, decoder depths $[2, 2]$, attention heads $[2, 4, 8]$, MLP ratio 4.0, dropout 0.1, and drop-path rate 0.1.
We train AdapToPASS-Swift for 400 epochs using AdamW with learning rate 2×10^{-4}, weight decay 10^{-2}, betas $(0.9, 0.999)$, cosine decay, minimum learning rate 10^{-6}, and 5 warmup epochs. Mixed precision is enabled, gradient clipping is set to 1.0, and cross-entropy is used as the training loss. The training batch sizes is 4. RGB inputs are normalized with mean 0.5 and standard deviation 0.225, and no data augmentation is used.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions of AdapToPASS, including ambiguity-aware spherical attention, bifocal spherical representation, SDF-based boundary supervision, and evaluation under unseen spherical transformations. The claims are supported by quantitative results on Stanford2D3D and WildPASS, ablation studies, and robustness evaluations across multiple transformation types and difficulty levels.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We include a dedicated discussion of failure cases (Sec 4.2). Although AdapToPASS consistently outperforms prior methods under unseen spherical transformations, its performance still degrades under extremely severe perturbations. As future work, we look forward to integrating stronger spherical representations, such as spherical harmonics and equivariant geometric priors, to further improve robustness under extreme spherical transformations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper should be numbered and cross-referenced.

   Guidelines:

   - The answer [N/A] means that answerNA
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the necessary methodological details and experimental settings required to reproduce the main results (Sec. 3 and Sec. 4), including implementation details, training setup, evaluation protocol, datasets, and additional configuration details in Appendix A and Appendix B. These details are intended to make the experimental setup clear and reproducible.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Click here to visit the project page.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We specify the training and test details in Sec. 4, including the dataset splits, evaluation protocol, and implementation settings. Additional experimental details and configuration settings are provided in the Appendix B to ensure completeness and reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report statistical significance through accuracy and mIoU scores in the relevant figures, where applicable, to indicate the variability and reliability of the experimental results. These error bars help support the robustness of the reported performance trends across the evaluated settings.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide the required compute resource details in Sec 4, including the hardware configuration and relevant training settings needed to reproduce the experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss the broader impacts of this work through the introduction, methodological motivation, and future work. The social impact is also discussed in Appendix C.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: We properly cite and acknowledge the existing assets used in this work, including public datasets, baseline methods, and any external code or models used for comparison. We use only the official open-source implementations released by the original authors whenever available. This ensures that all comparisons are based on properly licensed, publicly accessible, and reproducible assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We introduce new model components and experimental assets in this paper, and document them through the method section, experimental setup, and appendix.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.