Single Frame Semantic Segmentation Using Multi-Modal Spherical Images

Suresh Guttikonda   Jason Rambach
German Research Center for Artificial Intelligence (DFKI)
{suresh.guttikonda, jason.rambach}@dfki.de

Abstract

In recent years, the research community has shown a lot of interest to panoramic images that offer a 360° *directional perspective. Multiple data modalities can be fed, and complimentary characteristics can be utilized for more robust and rich scene interpretation based on semantic segmentation, to fully realize the potential. Existing research, however, mostly concentrated on pinhole RGB-X semantic segmentation. In this study, we propose a transformer-based cross-modal fusion architecture to bridge the gap between multi-modal fusion and omnidirectional scene perception. We employ distortion-aware modules to address extreme object deformations and panorama distortions that result from equirectangular representation. Additionally, we conduct cross-modal interactions for feature rectification and information exchange before merging the features in order to communicate long-range contexts for bi-modal and tri-modal feature streams. In thorough tests using combinations of four different modality types in three indoor panoramic-view datasets, our technique achieved state-of-the-art mIoU performance:* 60.60% *on Stanford2D3DS [2] (RGB-HHA),* 71.97% *on Structured3D [44] (RGB-D-N), and* 35.92% *on Matterport3D [5] (RGB-D).* [1]

1. Introduction

With the increased availability of affordable commercial 3D sensing devices, in recent years, researchers are more interested in working with omnidirectional images, also often referred to as 360°, panoramic, or spherical images. In contrast to pinhole cameras, the captured spherical images provide an ultra-wide 360° × 180° field-of-view (FoV) allowing for the capture of more detailed spatial information of the entire scene from a single frame [14, 43]. Practical applications of such immersive and complete view perception include holistic and dense visual scene understanding [1], augmented- and virtual reality (AR/VR) [26, 37], autonomous driving [11], and robot navigation [6].

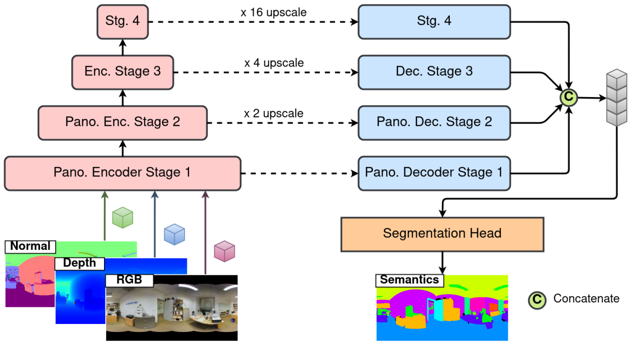
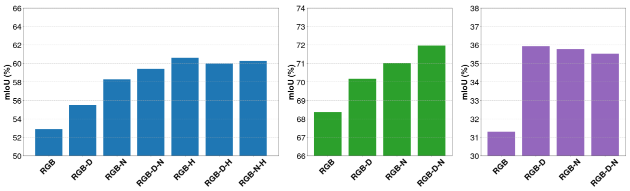


Figure 1. Overview of our multi-modal panoramic segmentation architecture. The inputs are an combination of **RGB**, **D**epth, and **N**ormals.

Generally, spherical images are represented using equirectangular projection (ERP) [38] or cubemap projection (CP) [31], which introduces additional challenges like scene discontinuities, large image distortions, object deformations, and lack of open-source datasets with diverse real-world scenarios. While extensive research has been conducted on pinhole based learning methods [4, 22, 24, 34, 35], approaches tailored for processing ultra-wide panoramic images and inherently accounting for spherical deformations remain ongoing research. Furthermore, the scarcity of labeled data, in indoor and outdoor scenarios, required for model training with panoramic images has slowed down the progress in this domain.

While previous panorama segmentation techniques have attained state-of-the-art performance for **RGB**-only images, they do not take advantage of the complementary modalities to develop discriminative features in situations when it is difficult to discriminate only based on texture information. With comprehensive cross-modal interactions for **RGB-X** modality [22], our work expands the current Trans4PASS+ [41] methodology for multi-modal panoramic semantic segmentation. For the Stanford2D3DS [2] dataset, we evaluate on 4 distinct multi-modal semantic segmentation tasks, including **RGB**, **RGB-D**epth, **RGB-N**ormal, and **RGB-H**HA, and we reach a

[1] Code will be made publicly available at https://github.com/sguttikon/SFSS-MMSI.



(a) 2D3DS [2] (b) Struct3D [44] (c) Mp3D [5]

Figure 2. Our cross-modal panoramic segmentation results with **RGB**, **D**epth, **N**ormals, and **H**HA combinations from Stanford2D3DS (*left*), Structure3D (*middle*) and Matterport3D (*right*) datasets.

state-of-the-art 60.60% with **RGB**-**H**HA semantic segmentation. We proposed a tri-modal fusion architecture and achieved top mIoU of 75.86% on Structure3D [44] (RGB-D-N) and 39.26% on Matterport3D [5] (RGB-D-N) for situations when HHA[2] is not accessible. The performance of our system on the aforementioned indoor panoramic-view datasets is shown in Fig. 2.

In summary, we provide the following contributions:

1. We investigate multi-modal panoramic semantic segmentation in four types of sensory data combinations for the first time.

2. We explore the multi-modal fusion paradigm in this study and introduce the tri-modal paradigm with cross-modal interactions for exploring texture, depth, and geometry information in panoramas.

3. On three indoor panoramic datasets that include RGB, Depth, Normal, and HHA sensor data combinations, our technique provides state-of-the-art performance.

2. Related Work

Semantic segmentation An encoder-decoder paradigm with two stages is typically used in existing semantic segmentation designs [3, 8]. A backbone *encoder* module [15, 17, 36] creates a series of feature maps in the earlier stage in order to capture high-level semantic data. Later, a *decoder* module gradually extracts the spatial data from the feature maps. Recent research has focused on replacing convolutional backbones with transformer-based ones in light of the success of vision transformer (ViT) in imagine classification [12]. Early studies mostly concentrated on the Transformer encoder design [9, 23, 33, 45], while later study avoided sophisticated decoders in favor of a lightweight All-MLP architecture [35], which produced results with improved efficiency, accuracy, and robustness.

Panoramic segmentation Early methods for interpreting a picture holistically centered on using perspective image-based models in conjunction with distorted-mitigated wide-field of view images. A distortion-mitigated locally-planar image grid tangents to a subdivided icosahedron is Eder *et al*. [13] novel proposal for a tangent image spherical representation. Lee *et al*. [21], on the other hand, uses a spherical polyhedron to symbolize comparable omnidirectional perspectives. Recent studies [25], however, use distortion-aware modules in the network architecture to directly operate on equirectangular representation. Sun *et al*. [30] suggests a discrete transformation for predicting dense features after an effective height compression module for latent feature representation. To improve the receptive field and learn the distortion distribution beforehand, Zheng *et al*. [46] combines the complimentary horizontal and vertical representation in the same line of research. In an encoder-decoder framework, Shen *et al*. [28] introduces a brand-new panoramic transformer block to take the place of the conventional block. Modern panoramic distortion-aware and deformable modules [10] have been added to the state-of-the-art UNet [27] and SegFormer [35] segmentation architectures to improve their performance in the spherical domain [14, 25, 40, 41].

Multimodal semantic segmentation Fusion strategies leverage the advantages of several data sources and show notable performance improvements for image-based semantic segmentation [7, 18]. The key contributions for comprehending **RGB-D** scenes concentrated on: 1) creating new layers or operators based on the geometric properties of **RGB-D** data [4, 7, 32], and 2) creating specialized architectures for combining the complimentary data streams in various stages [18, 20, 28, 30]. When modalities other than depth maps are employed, these approaches perform less well because they were created exclusively for **RGB-D** modality [42]. Recent studies have concentrated on establishing unique fusion algorithms for **RGB-X** semantic segmentation that are adaptable across various sensing modality combinations [22, 34, 39]. In the omnidirectional realm, however, the integration of several modalities with cross-modal interactions is still an unresolved issue. The main issue in this scenario is to recognize the distorted and deformed geometric structures in the ultra-wide 360-degree images while taking advantage of a variety of comprehensive complementing information. To jointly use the many sources of information from **RGB**, **D**epth, and **N**ormals equirectangular images, we propose our framework, which makes use of cross-modal interactions and panoramic perception abilities.

3. Methodology

Section 3.1 provides a summary of the framework we propose for panoramic multi-modal semantic segmentation.

[2] **H**orizontal disparity, **H**eight above ground, and normal **A**ngle to the vertical axis [16]

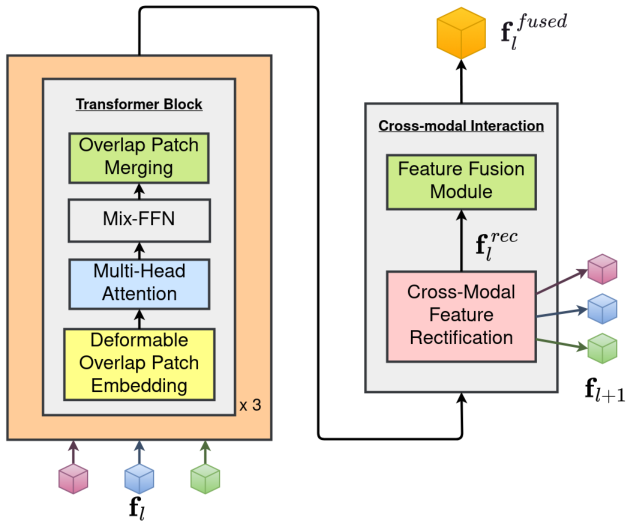


Figure 3. *Panoramic encoder stage* to extract **R**GB, **D**epth, and **N**ormals features.

Although our framework may be used for bi-modal and tri-modal input scenarios,for simplicity, we explain only the *encoder* and *decoder* architectures design for cross-modal (**RGB**-**D**epth-**N**ormals) panorama segmentation in Sec. 3.2 and Sec. 3.3, respectively. Our design is based on Trans4PASS+ [41] and uses an extension of CMX [22] for ternary modal streams feature extraction and fusion to learn object deformations and panoramic image distortions. We adopt a notation $\mathbf{f}$ to represent multi-modal feature maps, *i.e.* $\mathbf{f} \in \{\mathbf{f}_{rgb}, \mathbf{f}_{depth}, \mathbf{f}_{normal}\}$, in order to keep the notation simple and avoid the l notation for inputs and outputs to network modules in the l-th encoder-decoder stage.

3.1. Framework Overview

In accordance with Xie *et al.* [35], we proposed the multi-modal panoramic segmentation architecture depicted in Fig. 1. The $H \times W \times 3$ input image is first separated into patches. We provide panoramic hierarchical encoder stages to address the severe distortions in panoramas while allowing cross-modal interactions between **R**GB-**D**epth-**N**ormals patch features, as described in Sec. 3.2. The encoder uses these patches as input to produce multi-level features at resolutions of $\{1/4, 1/8, 1/16, 1/32\}$ of the original image. Finally, our panoramic decoder (refer Sec. 3.3) receives these multi-level features in order to predict the segmentation mask at a $H \times W \times N_{class}$ resolution, where N_{class} is the number of object categories.

3.2. Panoramic Hierarchical Encoding

Each stage of our encoding process for extracting hierarchical characteristics is specifically designed and optimized for semantic segmentation. Figure 3 illustrates how our architecture incorporates recently proposed Cross-modal Feature Rectification (CM-FRM) and Feature Fusion (FFM) modules [22] as well as Deformable Patch Embeddings (DPE) module [40] to deal with the severe distortions in **R**GB, **D**epth, and **N**ormals panoramas caused by equirectangular representation.

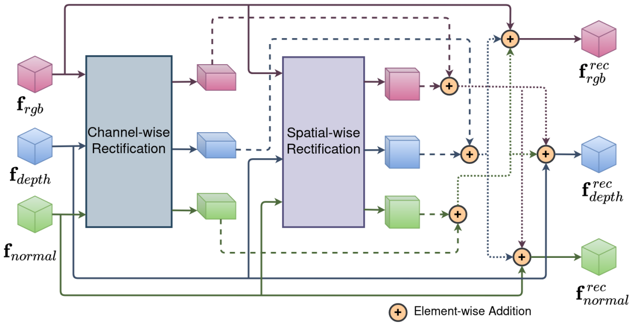

$\mathbf{f}_{rgb}$
$\mathbf{f}_{depth}$
$\mathbf{f}_{normal}$
Channel-wise Rectification
Spatial-wise Rectification
$\mathbf{f}_{rgb}^{rec}$
$\mathbf{f}_{depth}^{rec}$
$\mathbf{f}_{normal}^{rec}$
Element-wise Addition

Figure 4. *Cross-modal feature rectification module* to calibrate **R**GB, **D**epth, and **N**ormals features.

Deformable patch embedding A typical Patch Embeddings (PE) module [12, 35] divides an input image or feature map of size $\mathbf{f} \in \mathbb{R}^{H \times W \times C_{in}}$ into a flattened 2D patch sequence of shape $s \times s$ each. In this patch, the position offset with respect to a location (i, j) is defined as $\mathbf{\Delta}_{(i,j)} \in \left[\frac{-s}{2}, \frac{s}{2}\right] \times \left[\frac{-s}{2}, \frac{s}{2}\right]$, where $(i, j) \in [1, s]$. However, these fixed sample points fail to learn deformation-aware features and do not respect object shape distortions. To learn a data-dependent offset, we deploy a Deformable Patch Embeddings (DPE) module that was proposed by Zhang *et al.* [40]. We formulate Eq. (1), using the deformable convolution operation $g(.)$ [10] with a hyperparameter of $r = 4$.

$$\mathbf{\Delta}_{(i,j)}^{DPE} = \begin{bmatrix} min(max(-\frac{H}{r}, g(\mathbf{f})_{(i,j)}), \frac{H}{r}) \\ min(max(-\frac{W}{r}, g(\mathbf{f})_{(i,j)}), \frac{W}{r}) \end{bmatrix} \tag{1}$$

Cross-modal feature rectification Measurements that are noisy are frequently present in the data from various complementing sensor modalities. By utilizing features from a different modality, the noisy information can be filtered and calibrated. Regarding this, Liu *et al.* [22] present a novel Cross-Modal Feature Rectification Module (CM-FRM) to execute feature rectification between parallel streams at each stage, throughout feature extraction process. In our work, we expand this calibration scheme using ternary features from **R**GB, **D**epth, and **N**ormals panorama stream, as seen in Fig. 4. Our two-stage CM-FRM processes the input features channel- and spatial-wise to address noises and uncertainties in **R**GB-**D**epth-**N**ormals modalities, providing a comprehensive calibration for improved multi-modal feature extraction and interaction. While the spatial-wise rectification stage focuses on local calibration, the channel-wise rectification stage is

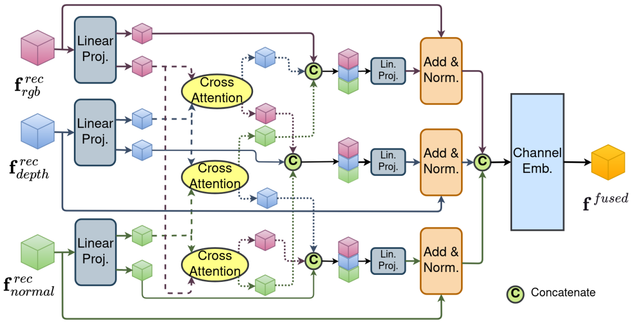


Figure 5. *Cross-modal feature fusion module* to fuse **RGB**, **D**epth, and **N**ormals features.

more concerned with global calibrations. Hyperparameters $\lambda_c, \lambda_s = 0.5$ are utilized to rectify the noisy input multi-modal features as shown in Eq. (2) by using the channel $\mathbf{f}^{rec}_{channel}$ and spatial $\mathbf{f}^{rec}_{spatial}$ weights that have been obtained.

$$\mathbf{f}^{rec} = \mathbf{f} + \lambda_c \mathbf{f}^{rec}_{channel} + \lambda_s \mathbf{f}^{rec}_{spatial} \tag{2}$$

Cross-modal feature fusion To improve information interaction and combine the features into a single feature map the rectified multi-modal feature maps $\mathbf{f}^{rec}$ are passed through a two-stage Feature Fusion Module (FFM) at the end of each encoder stage. As seen in Fig. 5, we use a ternary multi-head cross-attention mechanism to expand Liu *et al.* [22] information sharing stage by allowing for global information flow between the **RGB**, **D**epth, and **N**ormals modalities. In the fusion stage, a channel embedding [22] is utilized to combine ternary features to $\mathbf{f}^{fused}$ and passed through the decoding step for semantics prediction.

3.3. Panoramic Token Mixer Decoder

The vanilla All-MLP decoder employed in earlier works [35] lacked adaptivity to object deformations, which weakens the token mixing of panoramic data. A novel deformable token mixer, the DMLPv2, was proposed by Zhang *et al.* [41] and is demonstrated to be effective and lightweight for both spatial and channel-wise token mixing. We leverage the DMLPv2 token mixer approach at each l-th level of our framework, as depicted in Fig. 6, which is denoted as:

$$\hat{\mathbf{f}}_l = \mathbf{DPE}(\mathbf{f}_l^{fused}) \tag{3}$$

$$\hat{\mathbf{f}}_l = \mathbf{PX}(\hat{\mathbf{f}}_l) + \mathbf{CX}(\hat{\mathbf{f}}_l) \tag{4}$$

$$\hat{\mathbf{f}}_l = \mathbf{DMLP}(\hat{\mathbf{f}}_l) + \mathbf{CX}(\hat{\mathbf{f}}_l) \tag{5}$$

$$\mathbf{f}_l^{decoded} = \mathbf{UpSample}(\hat{\mathbf{f}}_l) \tag{6}$$

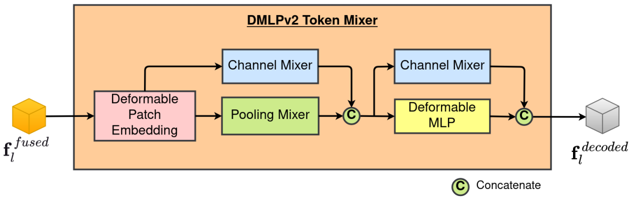


Figure 6. *Panoramic decoder stage* with fused features from **RGB**, **D**epth, and **N**ormals modalities.

The Channel Mixer (CX) of the DMLPv2 considers space-consistent yet channel-wise feature reweighting, strengthening the feature by emphasizing informative channels. Focusing on spatial-wise sampling using fixed and adaptive offsets, respectively, the Pooling Mixer (PX) and Deformable MLP (DMLP) are used in DMLPv2. The non-parametric Pooling Mixer (PX) is implemented by an average pooling operator. The adaptive data-dependent spatial offset $\mathbf{\Delta}^{DMLP}_{(i,j,c)}$ is predicted channel-wise.

Finally, to output the prediction for N_{class} semantics masks, the decoded features from the four steps are concatenated and given to a segmentation header module, depicted in Fig. 1.

4. Experiments

4.1. Datasets

For the purpose of evaluating our suggested cross-modal framework for interior settings, we use three multi-modal equirectangular semantic segmentation datasets. In each of our tests, we resize the input image to 512×1024, and then we compute evaluation metrics, such as Mean Region Intersection Over Union (mIoU), Pixel Accuracy (aAcc), and Mean Accuracy (mAcc), using the MMSegmentation IoU script[3].

Stanford2D3DS dataset [2] contains 1713 multi-modal equirectangular images with 13 object categories. We split the data from area_1 to area_6 for training and validation in a manner similar to Armeni *et al.* [2], using a 3-fold cross-validation scheme, and we give the mean values across the folds. Furthermore, the publicly accessible code[4] is used to compute the panoramic HHA [16] modality using the appropriate depth and camera parameters.

Structured3D dataset [44] offers 40 NYU-Depth-v2 [29] object categories, 196515 synthetic, multi-modal, equirectangular images with a variety of lighting setups. In line with Zheng *et al.* [44], we establish typical training, validation, and test splits as follows: scene_00000 to scene_02999 for training, scene_03000 to scene_03249 for validation, and scene_03250 to scene_03499 for testing. For

[3] https://mmsegmentation.readthedocs.io/en/0.x/

[4] https://github.com/charlesCXK/Depth2HHA-python

all of the tests we conduct, we use rendered raw lighting images with full furniture arrangements.

Matterport3D dataset [5] The 10800 panoramic views in the Matterport3D [5] collection are represented by 18 viewpoints per image frame, necessitating an explicit conversion to an equirectangular format. Second, the associated semantic annotations are spread among four files (xxx.house, xxx.ply, xxx.fsegs.json, and xxx.semseg.json). We employ the open-source matterport_utils[5] code for post-processing, where the *mpview* script is used to produce annotation images and the *preparepano* script is used to stitch the 18 images that were taken into a 360-degree panorama. For our trials using the 40 object categories, we created own training, validation, and test splits, refer to appendix.

4.2. Implementation Details

With an initial learning rate of 6e-5 programmed by the poly strategy with power 0.9 over the training epochs, we train our models using a pre-trained SegFormer MiT-B2[6] RGB backbone on the RTXA6000 GPU. For Stanford2D3DS [2], Structured3D [44], and Matterport3D [5] experiments, there are 200 training epochs, 50, and 100 respectively. The optimizer AdamW [19] is employed with the following parameters: batch size 4, epsilon 1e-8, weight decay 1e-2, and betas $(0.9, 0.999)$. Random horizontal flipping, random scaling to scales of $\{0.5, 0.75, 1, 1.25, 1.5, 1.75\}$, and random cropping to 512×512 are added for image argumentations. Deformable Patch Embedding module (DPE), refer to Sec. 3.2, is used for the panoramic encoder stage-1 and a conventional Overlapping Patch Embedding (OPE) module [35], for the other stages of our framework. More specific settings are described in detail in the appendix.

We conducted our tests for the following fusion configurations: **RGB**-only, **RGB**-**D**epth, **RGB**-**N**ormal, **RGB**-**H**HA, and **RGB**-**D**epth-**N**ormal, **RGB**-**D**epth-**H**HA, and **RGB**-**N**ormal-**H**HA. In our tests, we only use pathways and modules in our encoding-decoding stages and skip any unnecessary parts of our framework based on these combinations. For example, in the CM-FRM and FFM modules discussed in Sec. 3.2, we employ bi-directional features for cross-modal interactions for the **RGB**-**D**epth scenario, whereas for the **RGB**-**D**epth-**N**ormal situation, we use routes that lead to tri-directional interactions across the features.

4.3. Experiment Results and Analysis

We carry out comprehensive tests on multimodal segmentation datasets for indoor settings to demonstrate the effectiveness of our proposed architecture of cross-modal fusion using panoramas. We employ the aforementioned training epochs, random crop-size, and batch size variables to compare our method against the current state-of-the-art approaches Trans4PASS+ [41], HoHoNet [30], PanoFormer [28], CMNeXt [39], and TokenFusion [34]. For a detailed description of their implementation, see the corresponding works. While all other approaches have been reproduced using the conditions of our experiment, the CBFC [46] and Tangent [13] results described here are from the related original paper. In Figure 2, Figure 7 and Figure 8, as well as in Table 1 and Table 2, are visualizations of the quantitative results and comparisons to the state-of-the-art.

Method	**Modal**	**3-fold Val.**	
		mIoU (%)	**mAcc (%)**
Trans4PASS+ [41]	RGB	52.04	63.98
HoHoNet [30]		51.99	62.97
PanoFormer [28]		52.35	64.31
CBFC [46]		52.20	65.60
Tangent [13]		45.60	65.20
OURS		52.87	63.96
HoHoNet [30]	RGB-D	56.73	68.23
PanoFormer [28]		57.03	68.08
CBFC [46]		56.70	70.80
Tangent [13]		52.50	70.10
OURS		55.49	68.57
OURS	RGB-N	58.24	68.79
	RGB-H	**60.60**	**70.68**
	RGB-D-N	59.43	69.03
	RGB-D-H	59.99	70.44
	RGB-N-H	60.24	70.61

Table 1. Results on Stanford2D3DS [2].

Results on Stanford2D3DS Table 1 presents the thorough comparisons between our method and other current panoramic methods. Overall, our method delivers cutting-edge performance in the merging of complementary modalities for semantic segmentation. Our method produces results that are comparable to those of existing methods [13,28,30,46] when used with RGB-Depth panoramas, and it further improved the results when **RGB**, **D**epth, **N**ormals, and **H**HA combinations were combined. With **RGB**-**H**HA image-based fusion, the highest mIoU was reached at 60.60%. By utilizing the complementary geometric, disparity, and textural information, the mIoU metric increased from **RGB**-only to gradually fusing **D**epth and **N**ormals, $52.87\% \rightarrow 55.49\% \rightarrow 59.43\%$.

Results on Structured3D We further test Structured3D using simply **RGB**, **D**epth, and **N**ormals, as seen in Table 2. On the validation and test data splits, our **RGB**-only model performs at the cutting edge at 71.94% and 68.34%,

[5] https://github.com/atlantis-ar/matterport_utils

[6] https://github.com/huaaaliu/RGBX_Semantic_Segmentation

Method	**Modal**	**Structured3D**		**Matterport3D**	
		Validation mIoU (%)	**Test mIoU (%)**	**Validation mIoU (%)**	**Test mIoU (%)**
Trans4PASS+ [41]	RGB	66.74	66.90	33.43	29.19
HoHoNet [30]		66.09	64.41	31.91	29.33
PanoFormer [28]		55.57	54.87	30.04	26.87
OURS		71.94	68.34	35.15	31.30
HoHoNet [30]	RGB-D	69.51	66.99	35.36	32.02
PanoFormer [28]		60.98	59.27	33.99	31.23
OURS		73.78	70.17	39.19	**35.92**
OURS	RGB-N	74.38	71.00	38.91	35.77
	RGB-D-N	**75.86**	**71.97**	**39.26**	35.52

Table 2. Results on Structured3D [44] and Matterport3D [5] datasets.

respectively. Additionally, by combining depth and normals data, we were able to outperform benchmark results for (validation, test) by $(+1.84, +1.83)$ for **RGB-D**epth, $(+2.44, +2.66)$ for **RGB-N**ormals, and $(+3.92, 3.63)$ for **RGB-D**epth-**N**ormals fusion.

Results on Matterport3D Table 2 shows further trials using Matterport3D [5] with comparable **RGB**, **D**epth, and **N**ormals combinations in addition to the Structured3D [44] dataset. Our method outperforms the current panoramic techniques in this case for both **RGB**-only and **RGB-D**epth based semantic segmentation. Our validation and test pair mIoU metrics values for **RGB**-only and **RGB-D**epth, respectively, are $(35.15\%, 31.30\%)$ and $(39.19\%, 35.92\%)$, respectively, when compared to the benchmark. However, we discovered that the combination of the multi-modal fusion with normals did not result in the expected improvement in performance, as demonstrated in other tests, $(38.91\%, 35.92\%)$ for **RGB-N**ormal and $(39.26\%, 35.52\%)$ for **RGB-D**epth-**N**ormal. Our hypothesis is that the depth and normals data result in a limited amount of modal differences, and thus modal addition may be unnecessary.

4.4. Qualitative Analysis

The segmentation outcomes of panoramic techniques are shown in Fig. 7, which displays the findings from left to right and from top to bottom across several indoor datasets. Overall, our approach is able to take advantage of depth and geometry data as well as textures from **RGB**, **D**epth and **N**ormal modalities and correctly identify object semantics with a better level of accuracy, as indicated. While our baseline Trans4PASS+ [41] accurately classifies the book shelf, sofa, and chair in the first row, the architecture was unable to predict the exact geometrical shapes. Using depth information, PanoFormer [28] and HoHoNet [30] were able to estimate the exact geometry of the chair and bookshelf, however, former method incorrectly guessed the object class of the sofa. The third row findings of the **RGB**-only and **RGB-D**epth based techniques show a similar trend. When compared to current state-of-the-art baselines, our method consistently predicted geometric shapes that were considerably clearer and had precise object semantics in these situations. The approach can even handle thin structures like the neck of a guitar and items on a dining table, as shown in the second row.

The qualitative results of different Stanford2D3DS [2] multi-modal combinations, including **RGB**-only, **RGB-D**epth, **RGB-N**ormal, **RGB-HHA**, and **RGB-D**epth-**N**ormal, are shown in Fig. 8 using our paradigm. While in the scenarios shown in Fig. 8 (a) and Fig. 8 (b), using complementary data from other modalities is advantageous, this may not always be the case when the model cannot tell the difference between the distorted door and the wall (Fig. 8 (c)), or the distorted door and the bookshelf (Fig. 8 (d)). We hypothesize that these failed cases happened as a result of the scene objects' ambiguity, which makes it difficult to distinguish using any of the accessible modalities.

4.5. Ablation Studies

In the context of panoramic semantic segmentation, we investigated the state-of-the-art fusion architectures CMX [22], CMNeXt [39], and TokenFusion [34]. Our architecture, which was expanded to include a tri-modal panoramas scenario, is inspired on CMX [22]. In order to address panorama distortions, Deformable Patch Embeddings (DPE) modules, which are detailed in Sec. 3.2, are added to these encoder's backbone. The stages of the panorama decoder, as defined in Sec. 3.3, have not changed. We employ two versions of CMNeXt [39], one with and one without a Self-Query Hub (SQ-Hub), with the former version demonstrating the ability to handle up to 81 modalities with minimal overhead and processing demands. Furthermore, it is expected that SQ-Hub will soft-select informative features while remaining robust to sensor failure.

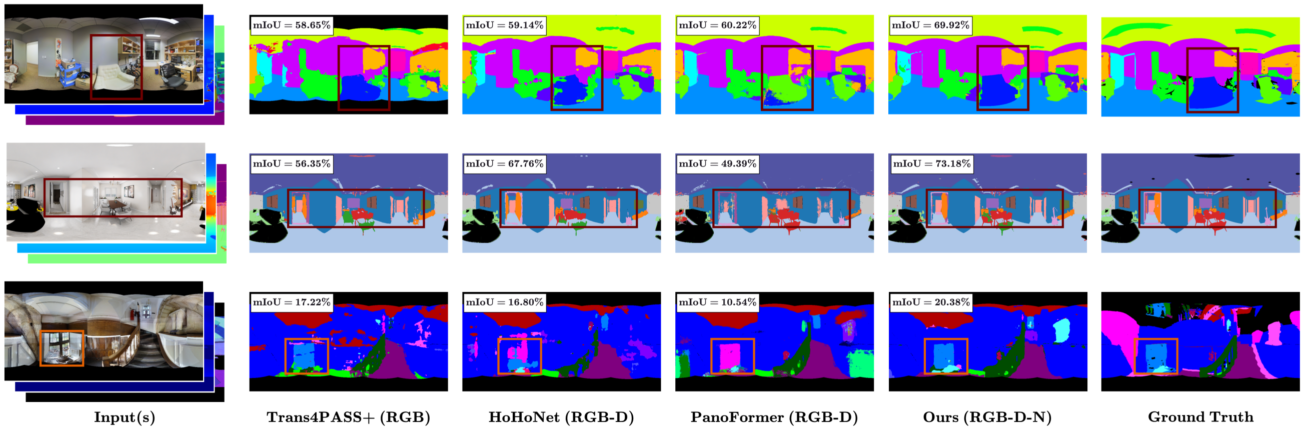


Figure 7. Results of multi-modal panoramic semantic segmentation for the **RGB**-only, **RGB**-**D**epth, and **RGB**-**D**epth-**N**ormals methods are visualized. For **RGB** segmentation, we use Trans4PASS+ [41] baseline, which employs the same SegFormer MiT-B2 backbone [35] with Deformable Patch Embeddings (DPE) and DMLPv2 decoder as ours, as detailed in Sec. 3.3. PanoFormer [28] uses a cutting-edge panoramic transformer-based architecture for **RGB**-**D**epth segmentation, while HoHoNet [30] is built on pre-trained ResNet-101 [17] in conjunction with a sophisticated horizon-to-dense module. Our strategy leverages **RGB**-**D**epth-**N**ormal fusion to improve performance by utilizing all available features.

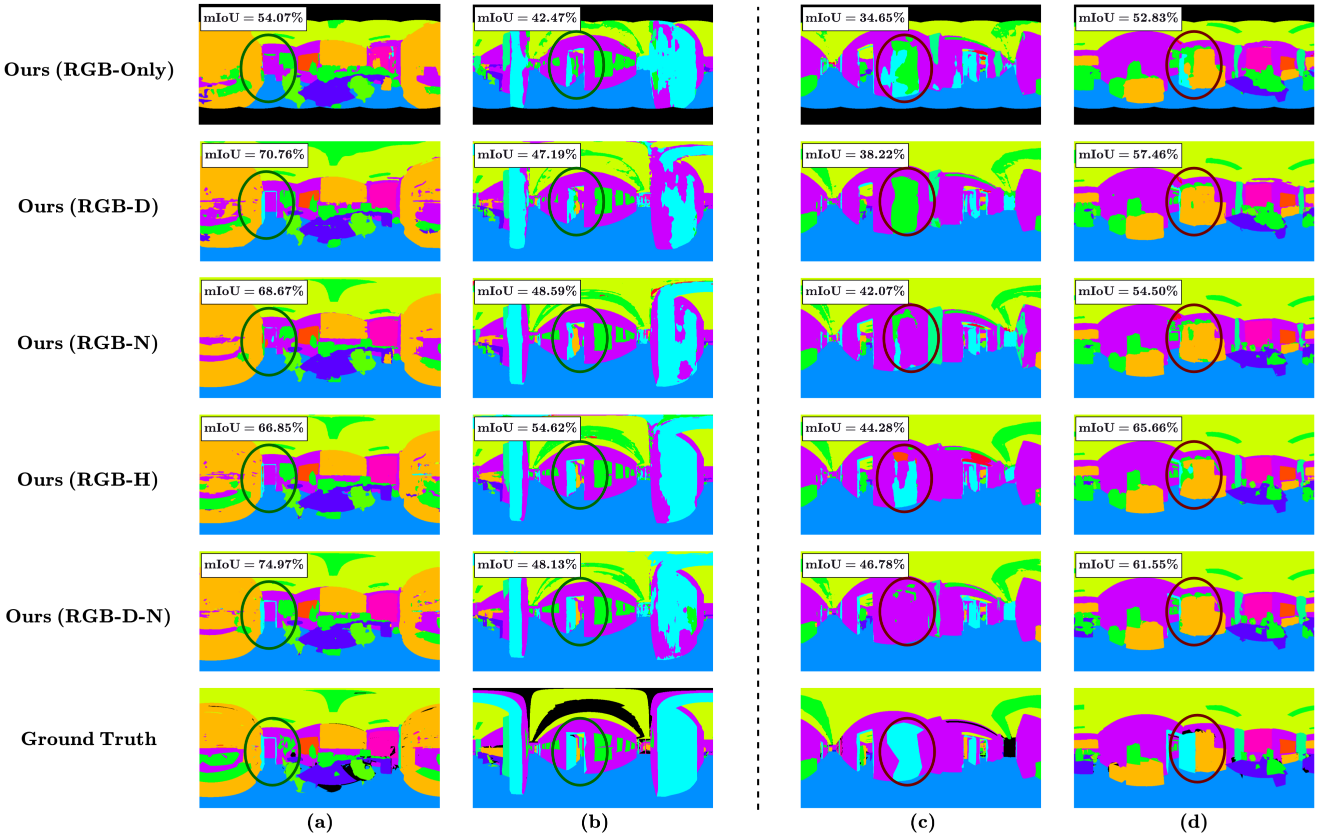


Figure 8. Visualization of semantic segmentation results for our framework using Stanford2D3DS [2] for **RGB**-only, **RGB**-**D**epth, **RGB**-**N**ormals, **RGB**-**H**HA, and **RGB**-**D**epth-**N**ormals (top-to-bottom) combinations. By utilizing complementary traits, our method was successful in identifying deformed and visually identical building structures like doors in columns (a) and (b). Under ambiguity, we were unable to differentiable between the distorted door and the wall or the deformed door and the bookcase in columns (c) and (d), respectively.

Method	Modal	Stanford2D3DS [2]		Structured3D [44]		Matterport3D [5]	
		mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)
OURS - TokenFusion [34]	RGB-D	58.88	68.57	62.58	70.54	**36.48**	49.30
OURS - CMNeXt (S) [39]		56.49	66.27	68.35	76.54	35.38	49.71
OURS - CMNeXt [39]		54.27	64.13	69.31	78.12	34.99	49.42
OURS		55.49	66.02	70.17	77.88	35.92	49.24
OURS - TokenFusion [34]	RGB-N	57.86	67.39	62.76	70.91	35.71	48.92
OURS - CMNeXt (S) [39]		53.61	63.26	68.47	76.82	33.10	46.32
OURS - CMNeXt [39]		50.47	60.83	68.62	76.99	33.80	47.02
OURS		58.24	68.79	71.00	78.68	35.77	**50.39**
OURS - TokenFusion [34]	RGB-H	59.06	68.07	–	–	–	–
OURS - CMNeXt (S) [39]		55.70	65.79	–	–	–	–
OURS - CMNeXt [39]		52.48	62.78	–	–	–	–
OURS		**60.60**	**70.68**	–	–	–	–
OURS - CMNeXt (S) [39]	RGB-D-H	57.62	67.80	–	–	–	–
OURS - CMNeXt [39]		54.54	64.22	–	–	–	–
OURS		59.99	70.44	–	–	–	–
OURS - CMNeXt (S) [39]	RGB-D-N	55.72	65.86	69.55	77.50	35.18	49.79
OURS - CMNeXt [39]		54.65	64.53	69.11	77.54	35.55	50.09
OURS		59.43	69.03	**71.97**	**79.67**	35.52	50.01
OURS - CMNeXt (S) [39]	RGB-N-H	55.45	65.24	–	–	–	–
OURS - CMNeXt [39]		52.50	62.19	–	–	–	–
OURS		60.24	70.62	–	–	–	–
OURS - CMNeXt (S) [39]	RGB-D-N-H	55.55	65.33	–	–	–	–
OURS - CMNeXt [39]		54.48	64.21	–	–	–	–

Table 3. An analysis of the various cross-modal fusion techniques applied to the encoder stages of our multi-modal panoramic architecture.

Table 3 compares { **RGB**-**D**epth, **RGB**-**N**ormals, and **RGB**-**H**HA } bi-modal fusion, { **RGB**-**D**epth-**N**ormal, **RGB**-**D**epth-**H**HA, and **RGB**-**N**ormal-**H**HA } tri-modal fusion, and { **RGB**-**D**epth-**N**ormal-**H**HA } quad-modal fusion. Overall, the CMX [22] technique we adopted had greater performance. Our methodology, which uses TokenFusion [34] for feature extraction and fusion, performs well on the Matterport3D [5] dataset, although it lags behind Stanford3D2DS [2] and Structured3D [44] by a wider margin. Thanks to Self-Query Hub (SQ-Hub), our approach to using encoded features from CMNeXt [39] performs comparably across datasets with fewer computational overload. However, in the majority of cases, in our panoramic trials, we have observed similar outcomes without SQ-Hub.

5. Conclusion

In this work, we revisit multi-modal semantic segmentation at the pixel level for a holistic scene understating. Through a cutting-edge panoramic encoder design, we present the framework with distortion awareness and cross-modal interactions. Our encoder learns severe object deformations and panoramic image distortions with equirectangular representations, and leverages feature interaction and feature fusion for cross-modal global reasoning in RGB-X panoramic segmentation. Our architecture produces superior performance on indoor panoramic benchmarks using RGB-Depth, RGB-Normal, and RGB-HHA combinations. Furthermore, we rebuild our cross-modal panoramic encoder to learn textual, disparity, and geometrical features using tri-modal (RGB-Depth-Normals) fusion, hence removing the requirement to compute HHA representations while maintaining the same performance. One major drawback of our method is that having two or more input streams active at once typically results in a large rise in complexity, refer to appendix. We'll look for techniques to combine multi-modal panoramas and 3D LiDAR data in the future with the least amount of processing effort possible.

Acknowledgement. This work was partially funded by the EU Horizon Europe Framework Program under grant agreement 101058236 (HumanTech).

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. *CoRR*, abs/2205.10468, 2022. 1

[2] Iro Armeni, Sasha Sax, Amir R. Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *CoRR*, abs/1702.01105, 2017. 1, 2, 4, 5, 6, 7, 8

[3] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(12):2481–2495, 2017. 2

[4] Jinming Cao, Hanchao Leng, Dani Lischinski, Danny Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor RGB-D semantic segmentation. In *ICCV*, pages 7068–7077. IEEE, 2021. 1, 2

[5] Angel X. Chang, Angela Dai, Thomas A. Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from RGB-D data in indoor environments. In *3DV*, pages 667–676. IEEE Computer Society, 2017. 1, 2, 5, 6, 8

[6] Devendra Singh Chaplot, Ruslan Salakhutdinov, Abhinav Gupta, and Saurabh Gupta. Neural topological SLAM for visual navigation. In *CVPR*, pages 12872–12881. Computer Vision Foundation / IEEE, 2020. 1

[7] Lin-Zhuo Chen, Zheng Lin, Ziqin Wang, Yong-Liang Yang, and Ming-Ming Cheng. Spatial information guided convolution for real-time RGBD semantic segmentation. *IEEE Trans. Image Process.*, 30:2313–2324, 2021. 2

[8] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV (7)*, volume 11211 of *Lecture Notes in Computer Science*, pages 833–851. Springer, 2018. 2

[9] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. In *NeurIPS*, pages 9355–9366, 2021. 2

[10] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773. IEEE Computer Society, 2017. 2, 3

[11] Grégoire Payen de La Garanderie, Amir Atapour Abarghouei, and Toby P. Breckon. Eliminating the blind spot: Adapting 3d object detection and monocular depth estimation to 360 ^\circ ∘ panoramic imagery. In *ECCV (13)*, volume 11217 of *Lecture Notes in Computer Science*, pages 812–830. Springer, 2018. 1

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*. OpenReview.net, 2021. 2, 3

[13] Marc Eder, Mykhailo Shvets, John Lim, and Jan-Michael Frahm. Tangent images for mitigating spherical distortion. In *CVPR*, pages 12423–12431. Computer Vision Foundation / IEEE, 2020. 2, 5

[14] Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, and José Jesús Guerrero. What's in my room? object recognition on indoor panoramic images. In *ICRA*, pages 567–573. IEEE, 2020. 1, 2

[15] Meng-Hao Guo, Cheng-Ze Lu, Qibin Hou, Zhengning Liu, Ming-Ming Cheng, and Shi-Min Hu. Segnext: Rethinking convolutional attention design for semantic segmentation. In *NeurIPS*, 2022. 2

[16] Saurabh Gupta, Ross B. Girshick, Pablo Andrés Arbeláez, and Jitendra Malik. Learning rich features from RGB-D images for object detection and segmentation. In *ECCV (7)*, volume 8695 of *Lecture Notes in Computer Science*, pages 345–360. Springer, 2014. 2, 4

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society, 2016. 2, 7

[18] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. ACNET: attention based network to exploit complementary features for RGBD semantic segmentation. In *ICIP*, pages 1440–1444. IEEE, 2019. 2

[19] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR (Poster)*, 2015. 5

[20] Seungyong Lee, Seong-Jin Park, and Ki-Sang Hong. Rdfnet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In *ICCV*, pages 4990–4999. IEEE Computer Society, 2017. 2

[21] Yeon Kun Lee, Jaeseok Jeong, Jong Seob Yun, Wonjune Cho, and Kuk-Jin Yoon. Spherephd: Applying cnns on a spherical polyhedron representation of 360deg images. In *CVPR*, pages 9181–9189. Computer Vision Foundation / IEEE, 2019. 2

[22] Huayao Liu, Jiaming Zhang, Kailun Yang, Xinxin Hu, and Rainer Stiefelhagen. CMX: cross-modal fusion for RGB-X semantic segmentation with transformers. *CoRR*, abs/2203.04838, 2022. 1, 2, 3, 4, 6, 8

[23] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 9992–10002. IEEE, 2021. 2

[24] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440. IEEE Computer Society, 2015. 1

[25] Semih Orhan and Yalin Bastanlar. Semantic segmentation of outdoor panoramic images. *Signal Image Video Process.*, 16(3):643–650, 2022. 2

[26] Minglang Qiao, Mai Xu, Zulin Wang, and Ali Borji. Viewport-dependent saliency prediction in 360° video. *IEEE Trans. Multim.*, 23:748–760, 2021. 1

[27] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI (3)*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 2

[28] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360$^{\circ }$ depth estimation. In *ECCV (1)*,

volume 13661 of *Lecture Notes in Computer Science*, pages 195–211. Springer, 2022. 2, 5, 6, 7

[29] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *ECCV (5)*, volume 7576 of *Lecture Notes in Computer Science*, pages 746–760. Springer, 2012. 4

[30] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *CVPR*, pages 2573–2582. Computer Vision Foundation / IEEE, 2021. 2, 5, 6, 7

[31] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *CVPR*, pages 459–468. Computer Vision Foundation / IEEE, 2020. 1

[32] Weiyue Wang and Ulrich Neumann. Depth-aware CNN for RGB-D segmentation. In *ECCV (11)*, volume 11215 of *Lecture Notes in Computer Science*, pages 144–161. Springer, 2018. 2

[33] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, pages 548–558. IEEE, 2021. 2

[34] Yikai Wang, Xinghao Chen, Lele Cao, Wenbing Huang, Fuchun Sun, and Yunhe Wang. Multimodal token fusion for vision transformers. In *CVPR*, pages 12176–12185. IEEE, 2022. 1, 2, 5, 6, 8

[35] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, pages 12077–12090, 2021. 1, 2, 3, 4, 5, 7

[36] Saining Xie, Ross B. Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated residual transformations for deep neural networks. In *CVPR*, pages 5987–5995. IEEE Computer Society, 2017. 2

[37] Mai Xu, Yuhang Song, Jianyi Wang, Minglang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2693–2708, 2019. 1

[38] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelhagen. Capturing omni-range context for omnidirectional segmentation. In *CVPR*, pages 1376–1386. Computer Vision Foundation / IEEE, 2021. 1

[39] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. *CoRR*, abs/2303.01480, 2023. 2, 5, 6, 8

[40] 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 *CVPR*, pages 16896–16906. IEEE, 2022. 2, 3

[41] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *CoRR*, abs/2207.11860, 2022. 1, 2, 3, 4, 5, 6, 7

[42] Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation. In *CVPR*, pages 2633–2642. Computer Vision Foundation / IEEE, 2021. 2

[43] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *ECCV (6)*, volume 8694 of *Lecture Notes in Computer Science*, pages 668–686. Springer, 2014. 1

[44] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *ECCV (9)*, volume 12354 of *Lecture Notes in Computer Science*, pages 519–535. Springer, 2020. 1, 2, 4, 5, 6, 8

[45] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, pages 6881–6890. Computer Vision Foundation / IEEE, 2021. 2

[46] Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, and Yao Zhao. Complementary bi-directional feature compression for indoor 360° semantic segmentation with self-distillation. In *WACV*, pages 4490–4499. IEEE, 2023. 2, 5