

Capturing Omni-Range Context for Omnidirectional Segmentation

Kailun Yang¹ Jiaming Zhang¹ Simon Reiß^{1,2} Xinxin Hu³ Rainer Stiefelhagen¹

¹CV:HCI Lab, Karlsruhe Institute of Technology ²Carl Zeiss AG ³Huawei Technologies

{kailun.yang, jiaming.zhang, rainer.stiefelhagen}@kit.edu

Abstract

Convolutional Networks (ConvNets) excel at semantic segmentation and have become a vital component for perception in autonomous driving. Enabling an all-encompassing view of street-scenes, omnidirectional cameras present themselves as a perfect fit in such systems. Most segmentation models for parsing urban environments operate on common, narrow Field of View (FoV) images. Transferring these models from the domain they were designed for to 360° perception, their performance drops dramatically, e.g., by an absolute 30.0% (mIoU) on established test-beds. To bridge the gap in terms of FoV and structural distribution between the imaging domains, we introduce Efficient Concurrent Attention Networks (ECANets), directly capturing the inherent long-range dependencies in omnidirectional imagery. In addition to the learned attention-based contextual priors that can stretch across 360° images, we upgrade model training by leveraging multi-source and omni-supervised learning, taking advantage of both: Densely labeled and unlabeled data originating from multiple datasets. To foster progress in panoramic image segmentation, we put forward and extensively evaluate models on Wild PANoramic Semantic Segmentation (WildPASS), a dataset designed to capture diverse scenes from all around the globe. Our novel model, training regimen and multi-source prediction fusion elevate the performance (mIoU) to new state-of-the-art results on the public PASS (60.2%) and the fresh WildPASS (69.0%) benchmarks.¹

1. Introduction

Convolutional Networks (ConvNets) reach striking performance on semantic segmentation [1, 34], a dense visual recognition task that aims at transforming an image into its underlying semantic regions. Particularly, segmenting images of road scenes automatically is of interest as it lies vital groundwork for scene understanding in an autonomous driving environment [26]. Most segmentation

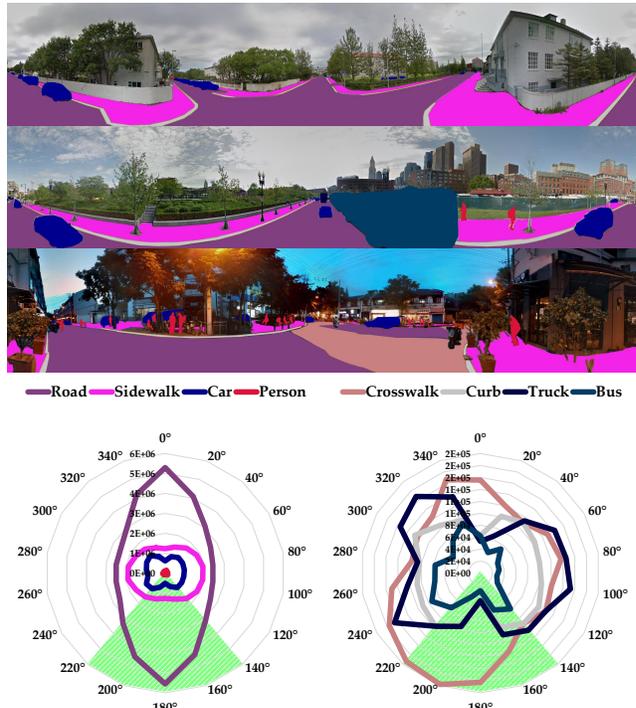


Figure 1. Top: Examples from our Wild PANoramic Semantic Segmentation (WildPASS) dataset; bottom: Class distribution of pixel-class associations as unfolded over the angular direction.

algorithms [8, 40, 73] are designed to work on pinhole-camera images whose Field of View (FoV) is rather narrow, capturing only a fraction of what is occurring on and aside the road. For a more holistic view on street-scenes, omnidirectional cameras are becoming ubiquitous in autonomous driving systems [71], as their capability of all-around sensing, gives rise to comprehensive 360° scene understanding. Yet, rather than designing models based on the new imaging modality, predominantly, due to lack of sufficient labeled data, narrow FoV-trained models are applied to yield a segmentation [7, 15, 59]. Brought about by the large mismatch in FoV and structural distribution between imaging domains, this practice causes significant performance degradation, even to a point of rendering the perception of surroundings completely unreliable [70].

Fig. 1 shows the distribution of semantic classes along

¹WildPASS: <https://github.com/elmino9ykl/WildPASS>

the viewing angle, where green-tinted regions indicate the section visible to a front-facing narrow FoV camera. It becomes apparent, that the positional priors in narrow- and omnidirectional images diverge severely, *e.g.*, the road class occupies much of the center portion in narrow FoV images, which, in 360° images, is far smaller at viewing angles perpendicular to the front-facing direction (180°). These observations suggest distinct contextual priors between the imaging techniques, which provide critical cues for semantic segmentation [78, 87]. Here, we address the challenging problem of leveraging inherent long-range contextual priors unique to omnidirectional images, that previous systems fail to harvest [69]. We lift the widely used, but computationally expensive non-local operations based on pixel-to-pixel correlations [17, 64] to the 360° FoV scenario in a computationally efficient fashion, *i.e.*, modeling correlations of each position to omni-range regions.

By design, non-local attention [64] is meant to aggregate associations from highly correlated positions like regions of homogeneous classes, which often lie in the horizontal direction at a similar height, *e.g.*, the sidewalks can be distributed across 360° but principally in the lower part of street-scenes. As preliminary investigation, we calculate the mean of the *Pearson correlation coefficient* $r(X_{dir})$ of the probability distribution over the classes in either the horizontal $r(X_{hor}) = 0.37$ or vertical $r(X_{ver}) = 0.16$ direction. In light of the substantially larger class-correlation of pixels observed horizontally, one can expect designing an attention module with emphasis in the large-FoV width-wise dimension to be advantageous, which reinforces the model’s discriminability while eliminating redundant computation on weakly-correlated pixels belonging to heterogeneous semantics. With this rationale, we propose *Efficient Concurrent Attention Networks (ECANets)* for capturing omni-range dependencies within wide-FoV imagery.

Aside from considering the inherent properties of 360° data in architecture design, training such models is not straightforward, as a lack of sufficient annotated panoramic images impedes optimization. The labeling process is extremely time- and labor-intensive, thus, only few surround-view datasets [52, 63, 76] are publicly accessible and none capture as diverse scenes as contemporary pinhole datasets [38, 58] do. This is why, to unlock the full potential of our novel ECANet architecture, we propose a *multi-source omni-supervised learning regimen* that intertwines training on unlabeled, full-sized panoramic images and densely labeled pinhole images. By means of data distillation [46] and our novel multi-source prediction fusion, we increase prediction certainty, ingrain long-range contextual priors while still benefiting from large-scale, narrow FoV segmentation datasets.

To facilitate progress in omnidirectional semantic image segmentation, we put forward the *Wild PANoramic Sema-*

tic Segmentation (WildPASS) dataset for evaluating models in unconstrained environments, featuring scenes collected from all around the globe, reflecting the perception challenge of autonomous driving in the real world. An exhaustive number of diverse experiments demonstrate the proposed ECANet model, learning- and fusion strategy make it possible to deploy highly efficient ConvNets for panoramic image segmentation, surpassing the state-of-the-art on the public PASS [69] and novel WildPASS datasets. On a glance, we deliver the following contributions:

- Rethink omnidirectional semantic segmentation from the context-aware perspective and propose *Efficient Concurrent Attention Networks*, capturing inherent long-range dependencies across the 360°.
- Introduce *multi-source omni-supervised learning*, integrating unlabeled panoramic images into training, empowering models to learn rich contextual priors.
- Present the diverse *WildPASS dataset*: Collected from 6 continents and 65 cities, enabling evaluation of panoramic semantic segmentation in the wild.
- Our methods surpass previous and set new state-of-the-art results on the PASS and WildPASS datasets.

2. Related Work

Context-aware semantic segmentation. Semantic segmentation has progressed exponentially since the conception of fully convolutional networks [34] and the groundwork laid by early encoder-decoder architectures [1, 39, 49]. Building atop classification networks [18, 22], DeepLab [4, 5], PSPNet [91], RefineNet [31] and DenseASPP [73] leverage pre-trained models and achieve notable performance improvements. Yet, a lot of progress is driven by sub-modules like dilated convolutions [81], large kernels [43], pyramid-, strip- and atrous spatial pyramid pooling [4, 19, 91] for capturing structured context.

A second branch of work takes advantage of recent channel/spatial attention mechanisms [20, 21, 62, 65] to exploit global context [29, 86], dimension-wise priors [8, 72] and cross-modal features [55, 67, 88]. More recently, inspired by non-local blocks in recognition tasks [64], DANet [17] and OCNet [83] adopt self-attention [60] to capture either associations between any pair of pixels/channels or dense object context, respectively. Further, point-wise attention is adaptively learned in [92]. This trend prompts a variety of attention modules [6, 24, 30, 32], as well as graph-based models [77, 89] and factorized variants [23, 51, 75]. With the nature of dense prediction in segmentation, the computational cost induced by combinatorics of pixel-pairs becomes infeasible very fast, which is amplified when working on pixel-rich, extra-wide panoramic images. In order to reduce these pair-wise computations, CCNet [25] operates in criss-cross paths, Axial-DeepLab [61] propagates along height- and width-axis sequentially, while [16, 53, 77, 94]

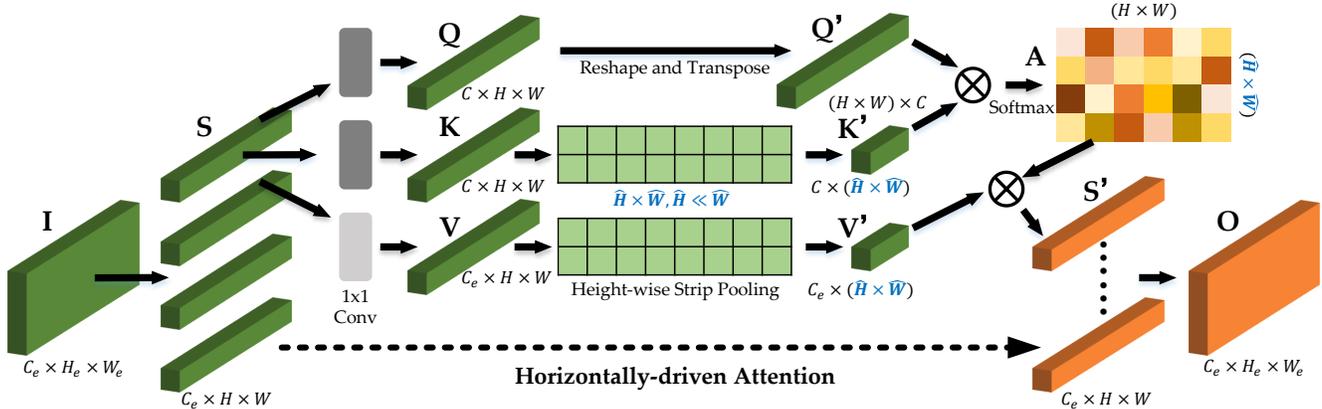


Figure 2. *Horizontal Segment Attention (HSA)* module. The omnidirectional image feature is divided into segments to model self-attention for capturing horizontal contextual information. To further reduce computation complexities and emphasize dependencies along the 360° width-wise direction, the association is measured between each pixel and gathering regions obtained with height-wise strip pooling.

act on key responding positions. Addressing the nature of omni-range contextual priors in panoramic images, we propose an efficient concurrent attention module. Unlike context-aware networks [19, 25] that only aggregate spatial dependency or refine pixel-level context, our attention concurrently highlights horizontally-driven dependencies and collects global contextual information for wide-FoV segmentation. While accuracy-oriented networks reach high performance, efficient networks like Fast-SCNN [45], CGNet [66] and ERFNet [48] aim to perform both swift and accurate segmentation. Further, a subset of compact ConvNets [66, 70, 79] utilize attention mechanisms to efficiently aggregate context. However, these methods do not consider inherent properties of omnidirectional imagery, and thus suffer from large accuracy degradation on such data. Our solution to solve this issue extends the line of efficient ConvNets and consistently elevates their performance and reliability when operating on panoramic images.

Omnidirectional semantic segmentation. As systems progressed towards 360° perception, early omnidirectional segmentation works were based on fisheye images [12, 74] or focused on indoor scenes [54, 56] with surround-view systems in street-scenes being largely based on a multitude of cameras [11, 37, 41, 47]. Motivated by the prospect of attaining 360° semantic perception based on a singular panoramic camera, recent works build semantic segmentation systems directly on top of this sensor-modality [2, 68, 85]. However, these methods rely on either expensive manually labeled or synthesized omnidirectional data, that does not capture the diversity nor the realism as seen in large-scale pinhole image collections with rich ontologies [38, 58]. With this sentiment in mind, we introduce the WildPASS dataset, comprising 360° images from over 60 cities and multiple continents, encouraging a more realistic assessment of panoramic segmentation performance. Yang *et al.* presented a framework [69, 70] for re-using models trained on pinhole images via style trans-

fer [93] and separating the panorama into multiple partitions for predictions, each resembling a narrow FoV image. While quite accurate, running multiple forward passes suffers from high latency [69] and disregards global context information within the panorama, as it is divided into patches. With our holistic view on panoramic image segmentation, we cut this computational burden by nearly a factor of four, extending the framework in [71] with a multi-source omni-supervised learning regimen that covers 360° imagery via data distillation [46] as well as enjoying the advantages of large, readily available, labeled pinhole image datasets. Further, making best use of the multi-source setting, we integrate a prediction fusion scheme, improving semantic certainty and generalization, as indicated by performance measures on both: PASS and WildPASS datasets.

3. Methodology

3.1. Efficient Concurrent Attention

To capture omni-range contextual priors in panoramic images, we propose *Efficient Concurrent Attention Networks (ECANets)*. Classical non-local methods [17, 64] generate large affinity maps to measure the association among each pixel-pair, producing a complexity in space of $\mathcal{O}((H_e \times W_e) \times (H_e \times W_e)) \sim \mathcal{O}(H_e^2 W_e^2)$. $H_e \times W_e$ denotes the spatial resolution of the feature map, therefore at the rather small scale of 128×64 , the complexity already amounts to 67,108,864 attention-related computations. This substantially increases training footprint and inference requirement, thus hindering memory-efficient learning and fast execution. With the assumption that wide-FoV panoramas contain rich contextual dependencies in the horizontal dimension, our ECANet features a *Horizontal Segment Attention (HSA)* module (Fig. 2) and a *Pyramidal Space Attention (PSA)* module (Fig. 3) for efficient context aggregation, simultaneously lessening the computational burden to a large extent as compared to the non-local base-

line [64]. As shown in Fig. 4, the attended-to feature maps are passed through PSA at different scales and then concatenated with HSA- and backbone feature maps. Finally, the resulting feature map is transformed by convolution- and upsampling layers to yield the semantic maps.

Horizontal segment attention module. The HSA module (Fig. 2) produces horizontally-driven feature maps for a given input $\mathbf{I} \in \mathbb{R}^{C_e \times H_e \times W_e}$ obtained from the backbone. For street-scenes, wide-FoV panoramas contain rich dependencies in the spatial dimension W , e.g., shared semantics between distant patches of sidewalk along the 360° can help in producing more consistent predictions. In this sentiment and with the aim of modeling self-attention to capture the wide FoV dependencies, the input feature is first divided into N segments ($N = 4$ in Fig. 2) along the H dimension. Each segment $\mathbf{S} \in \mathbb{R}^{C_e \times H \times W}$ has a smaller size with $(H, W) = (\frac{H_e}{N}, W_e)$, which is fed into convolution layers to generate feature maps of query \mathbf{Q} , key \mathbf{K} and value \mathbf{V} to carry out self-attention [60]. This reinforces the horizontal context aggregation and reduces the complexity to $\mathcal{O}(N \times (\frac{H_e}{N} \times W_e) \times (\frac{H_e}{N} \times W_e)) \sim \mathcal{O}(H_e^2 W_e^2 / N)$.

Further, advocating 360° region-level priors and to largely lessen computation burdens, we propose to apply a height-wise strip pooling on both the key feature $\mathbf{K} \in \mathbb{R}^{C \times H \times W}$ and value feature $\mathbf{V} \in \mathbb{R}^{C_e \times H \times W}$. The resulting size of this strip pooling is $\hat{H} \times \hat{W}$, where $\hat{H} \ll \hat{W}$ but $\hat{H} \neq 1$ to enable interactions between vertically adjacent patches. The attention map \mathbf{A} , is then computed from the reshaped query feature $\mathbf{Q}' \in \mathbb{R}^{(H \times W) \times C}$ and the new key feature $\mathbf{K}' \in \mathbb{R}^{C \times (\hat{H} \times \hat{W})}$:

$$\mathbf{Q}' \times \mathbf{K}' \rightarrow \mathbf{A} \in \mathbb{R}^{(H \times W) \times (\hat{H} \times \hat{W})} \quad (1)$$

In this manner, the association between each pixel and strip-pooled regions is ingrained in the module’s architecture, which directly enables learning relations along the horizontal axis, across the 360° . We apply the softmax function to the attention map, transpose it and attend to the value feature $\mathbf{V}' \in \mathbb{R}^{C_e \times (\hat{H} \times \hat{W})}$ as follows:

$$\mathbf{V}' \times \mathbf{A}^T \rightarrow \mathbf{S}' \in \mathbb{R}^{C_e \times (H \times W)} \quad (2)$$

Finally, all segments $\mathbf{S}' \in \mathbb{R}^{C_e \times H \times W}$ are concatenated along the vertical dimension H amounting to the horizontally-driven output $\mathbf{O} \in \mathbb{R}^{C_e \times H_e \times W_e}$. With this architectural design, we are able to greatly reduce the complexity to $\mathcal{O}(N \times (\frac{H_e}{N} \times W_e) \times (\hat{H} \times \hat{W})) \sim \mathcal{O}(H_e W_e \hat{H} \hat{W})$, as $\hat{H} \hat{W} \ll H_e W_e$. Further, our HSA module selectively aggregates context, as each pixel gathers information from its semantically-correlated regions in the horizontal direction.

Pyramidal space attention module. To capture global context, we introduce the PSA module (Fig. 3), relating each pixel to globally distributed regions using spatial pooling. We depict the pyramid scales of the spatial pooling in Fig. 4 with $(\hat{H}, \hat{W}) = \{(3, 3), (4, 4), (6, 6)\}$. In this manner, the complexity is $\sum_i \mathcal{O}(H_e W_e \hat{H}_i \hat{W}_i)$, a con-

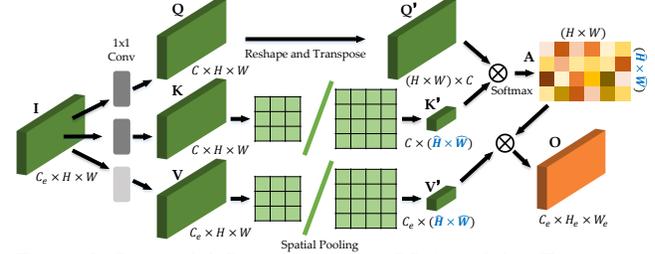


Figure 3. *Pyramidal Space Attention (PSA)* module. To capture global context in an efficient way, self-attention is modeled with the association measured between each pixel and globally distributed regions sampled using spatial pooling of pyramid scales.

siderable reduction as compared to the non-local baseline. The attended-to feature maps are concatenated to the HSA module output and the features produced by the backbone. Thereby, the proposed ECANet gathers both: wide-FoV horizontal semantics and global dependencies for enhanced omni-range reasoning.

3.2. Multi-source Omni-supervised Learning

In the following, we consider the training of panoramic segmentation networks and propose a better suited regimen.

Multi-source omni-supervision. We extend the design of [71] based on data distillation [46] and present a *multi-source omni-supervised learning regimen*. As shown in Fig. 4, we train efficient networks on both labeled pinhole- and unlabeled panoramic images in a multi-source setup. This adds diversity of FoV while simultaneously preventing the model from overfitting to pinhole images. While different datasets have heterogeneous class definitions, therefore hindering direct training by merging sources, we argue that the relationships encoded in their similar label hierarchies positively affect the generalization of feature representations. For example, street-scene datasets [9, 38, 58] have different definitions of road surfaces, but share similar class hierarchies that both comprise flat region and traffic object categories. As shown in Fig. 4, our approach is designed to enable learning with unequal sources in spite of their class-label conflicts. Precisely, to address the heterogeneity in semantic spaces, we append multiple output-heads to the efficient learner, each of which is a fully convolutional module with an upsampling layer for prediction in different label spaces (semantic spaces 1 and 2 for two sources, as illustrated in Fig. 4).

For unlabeled panoramas, we create annotations using a sophisticated architecture pre-trained on pinhole sets by means of the PASS pipeline [69], as such we leverage the correspondence of FoV between pinhole images and partitioned panoramas for accurate segmentation. The panorama is rotated along the horizontal direction based on its 360° wrap-around structure, and the predictions on differently rotated transformations are ensembled, producing the annotation. While time-consuming due to the complex PASS

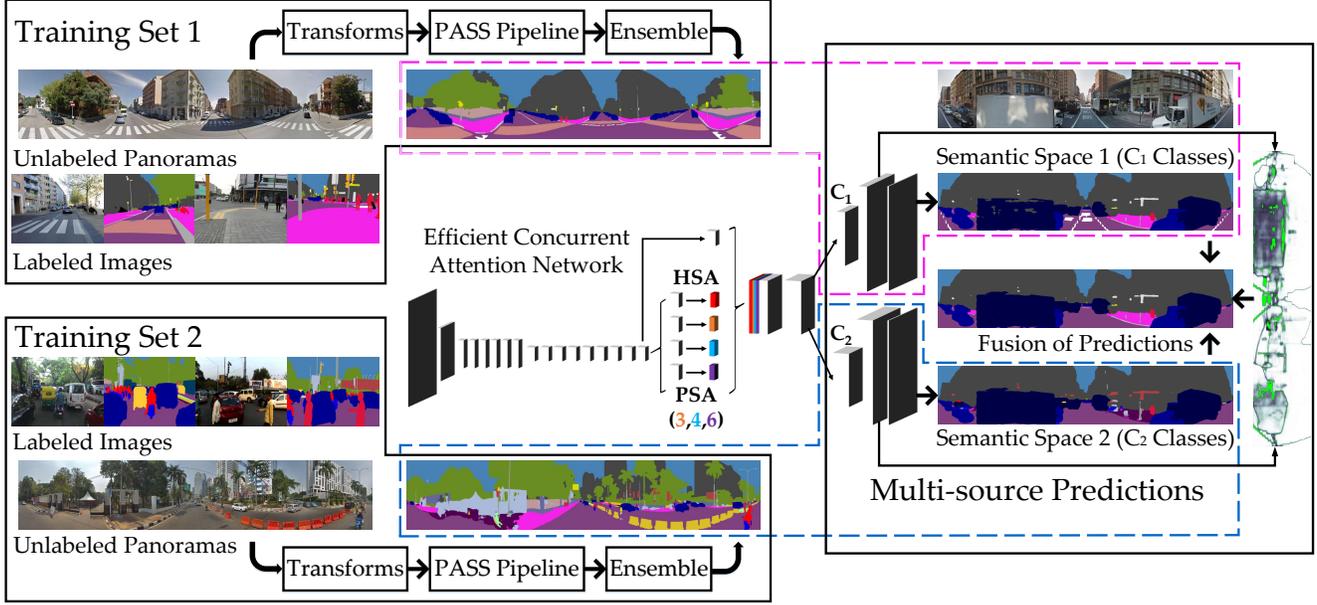


Figure 4. Diagram of the *multi-source omni-supervised learning regimen* for omnidirectional semantic segmentation. During training, the efficient concurrent attention network is trained on the union of labeled pinhole images and unlabeled panoramas in a multi-source manner. In the case of two domains, the network is enhanced with two heads for segmentation in heterogeneous label spaces. Annotations for the panoramas are automatically generated by seizing an ensemble of predictions on multiple transformations with the PASS pipeline [69]. During testing, the efficient produces two sets of semantic maps with classes C_1 and C_2 , which are fused in a post-processing step. On the right is an example of multi-space fusion, where green pixels on the uncertainty map denote areas with refined semantics (see the truck).

pipeline and ensembling, the resulting fine-grained, high quality annotations are computed in advance and can be leveraged for data distillation (Fig. 4). Unlike [71] that uses a battery of single-city unlabeled images, we distill knowledge via diverse panoramas from all around the globe, which will unfold as crucial detail for attaining generalization to open omnidirectional domains.

Multi-space fusion. In [71], while multi-space predictions are produced, only a single semantic map (*i.e.*, the Vistaspace) is used for the inferred omnidirectional perception. We assess this to be sub-optimal as it disregards the overlap in labels of the multi-space semantic spaces. To address this, we propose to fuse the multi-space predictions in a post-processing step. As such, we consider the intersecting classes of all semantic spaces C_j : $C_{IN} = \bigcap_{j=1}^J C_j$. Based on the unnormalized output \mathbf{z}_i^j for the i^{th} pixel as produced by the trained j^{th} output-head, the predicted class y_i^j is known as $y_i^j = \arg \max_{c \in C_j} \mathbf{z}_{i,c}^j$. However, considering other candidate heads, the prediction will be updated as:

$$y_i^j = \arg \max_{c \in C_{IN}} \mathbf{z}_{i,c}^{j*}, \quad (3)$$

if the prediction of the optimal head j^* is within the intersecting classes C_{IN} . To find the optimal output-head j^* , several strategies can be leveraged. A straightforward method based on the minimal uncertainty of the prediction is $j_{\text{var}}^* = \min_{j \in J} (\text{var}(\mathbf{z}_i^j))$, where uncertainty is modeled by the variance $\text{var}(\mathbf{z}_i^j) = (1/C_j) \sum_{c=1}^{C_j} (\mathbf{z}_{i,c}^j - \bar{\mathbf{z}}_i^j)^2$. An exemplary uncertainty map is visualized on the right side

of Fig. 4. Likewise, a common method for fusing predictions is to simply choose the maximum confidence based on the prediction probability: $j_{\text{conf}}^* = \max_{j \in J} \max \sigma(\mathbf{z}_{i,c}^j)$ with $\sigma(\cdot)$ denoting the softmax function. However, to select the optimal prediction, we make use of the calibrated confidence as given by the ratio of the highest and second highest probabilities: $j_{\text{cal}}^* = \max_{j \in J} \frac{\max_1 \mathbf{z}_{i,c}^j}{\max_2 \mathbf{z}_{i,c}^j}$. Compared to using all values in j_{var}^* or just the maximum one in j_{conf}^* , this fusion refines panoramic segmentation by considering the top-2 predictions, as shown in Fig. 4.

4. Experiments

4.1. WildPASS Dataset

To foster research and to facilitate credible numerical evaluation for panoramic scene segmentation, we introduce the *Wild Panoramic Semantic Segmentation* (WildPASS) dataset. Unlike mainstream large-scale datasets like Cityscapes [9] and BDD [80] that focus on urban scene understanding in Europe or North America, WildPASS embraces segmentation in the wild with images from all around the world based on Google Street View. As shown in Fig. 5, it includes unstructured and adverse road scenes like nighttime, as well as images taken from different viewpoints to paint a comprehensive picture of real-world challenges for navigational perception systems. Overall, we gather 2500 panoramas in 65 cities from all continents Asia, Europe,



Figure 5. Example images in the WildPASS dataset, including fine labeled and unlabeled panoramas collected from all around the world.

Africa, Oceania, North and South America (excluding only Antarctica). Following PASS [69] and RainyNight [13], 500 panoramas from 25 cities are precisely annotated for the most critical navigation-related classes. We extend the semantics defined for panoramic segmentation [69] to 8 classes: *Car*, *Road*, *Sidewalk*, *Crosswalk*, *Curb*, *Person*, *Truck* and *Bus*. The remaining 2000 panoramas from 40 different cities are collected for omni-supervised learning. All images have a $360^\circ \times 70^\circ$ FoV at 2048×400 . WildPASS thereby is the largest panoramic street-scene segmentation benchmark, exceeding other evaluation-oriented WildDash (211 annotations) [84], ISSAFE (313 annotations) [88] and PASS (400 annotations) [69] datasets.

4.2. Experimental Setup

Our experiments are built on the Mapillary Vistas [38] and IDD20K [58] pinhole datasets for a multi-source setting. These datasets are appealing as they feature diverse viewpoints and unstructured scenes inherent in real-world unconstrained surroundings. Both datasets contain 18000/2000 and 14027/2036 images for training/validation, respectively. Following [71], we use 25 classes from Vistas and the level-3 labels (26 classes) of IDD20K, 15 of which overlap. We use the 2000 unlabeled panoramas of WildPASS to create omnidirectional annotations with PSPNet50 [91] for data distillation. We test on public PASS (400 annular images) and WildPASS (500 panoramas) datasets and evaluate with the mean Intersection over Union (mIoU) metric. Facilitating fair comparisons with omnidirectional segmentation frameworks [69, 71], ECANet is implemented using the ERFNet [48] backbone pre-trained on ImageNet [50]. We use $N = 4$ segments and a strip pooling of $(\hat{H}, \hat{W}) = (2, 16)$ in HSA and spatial pooling of $(\hat{H}, \hat{W}) = \{(3, 3), (4, 4), (6, 6)\}$ in PSA. ECANet is trained under Adam optimization [28] with a Weight Decay of 2×10^{-4} and a Learning Rate of 1×10^{-4} that decreases exponentially over 200 epochs. We train with a batch-size of 12 and a resolution of 1024×512 . For multi-source training, each iteration comprises a forward and backward pass per dataset with standard cross-entropy loss throughout.

4.3. Results on PASS

As shown in Table 1, we first present the results on the public Panoramic Annular Semantic Segmentation (PASS) dataset [69], which represents an unseen test-bed to our

Network and Method		mIoU
Vistas-trained	Fast-SCNN [45]	28.5%
Vistas-trained	SegNet [1]	25.7%
Vistas-trained	DRNet (ResNet18) [82]	28.0%
Vistas-trained	PSPNet (ResNet50) [91]	41.4%
Vistas-trained	DenseASPP (DenseNet121) [73]	33.3%
Vistas-trained	DANet (ResNet50) [17]	38.9%
Style-transferred	ENet [42]	31.0%
Style-transferred	SQNet [57]	27.9%
Style-transferred	ERFNet [48]	34.3%
Style-transferred	LinkNet [3]	30.5%
Style-transferred	PSPNet (ResNet18) [91]	34.8%
Style-transferred	ICNet [90]	25.7%
Style-transferred	ESPNet [36]	24.7%
Style-transferred	BiSeNet [79]	27.7%
Style-transferred	EDANet [33]	30.5%
Style-transferred	CGNet [66]	30.4%
Style-transferred	SwiftNet [40]	37.4%
Style-transferred	SwiftNet [70]	38.2%
Vistas-trained	ERF-PSPNet	32.2%
Style-transferred	ERF-PSPNet	39.2%
PASS [69]	ERF-PSPNet	58.2%
OOSS [71]	ERF-PSPNet	47.9%
Our method	ERF-PSPNet	58.4%
Our method	ECANet	60.2%

Table 1. Accuracy analysis on the public PASS dataset [69].

models. Thereby, the evaluation paradigm expects a trained model that can generalize well in an open, previously unseen domain. A variety of networks has been benchmarked on the PASS dataset, including some trained on Mapillary Vistas and some style-transferred using CycleGAN [69, 70]. Overall, even though the networks under our multi-source omni-supervision have not seen any annular image from the PASS domain, they outperform all previous models by large margins. In comparison to the PASS method using ring-padding [10] for continuous segmentation and the single-pass OOSS method [71], our improvement is significant thanks to the diversity introduced in the data distillation. Moreover, our ECANet capturing omni-range dependencies, boosts mIoU to 60.2%, setting new state-of-the-art accuracy among 20 mainstream networks.

4.4. Results on WildPASS

Comparison with state-of-the-art. With the novel WildPASS dataset, we establish a benchmark for omnidirectional segmentation. The most accurate and efficiency-oriented architectures are trained on Mapillary Vistas and extensively evaluated on WildPASS. Table 2 shows their per-class accuracy as well as their computation complexity quantified by Multiply-Accumulate operations (MACs) and

Network	mIoU	Car	Road	Sidew.	Crossw.	Curb	Person	Truck	Bus	MACs	PARAMs
SegNet [1]	22.7%	57.2%	61.1%	18.2%	4.2%	14.1%	5.8%	9.6%	11.1%	398.3G	28.4M
PSPNet50 [91]	46.1%	80.0%	74.9%	51.7%	23.9%	31.4%	19.8%	38.9%	48.1%	403.0G	53.3M
DenseASPP [73]	33.2%	51.1%	69.1%	38.1%	16.4%	26.3%	8.7%	27.4%	28.4%	78.3G	8.3M
DANet [17]	47.2%	74.8%	72.2%	49.9%	28.9%	23.8%	25.4%	51.9%	50.6%	114.1G	47.4M
ENet [42]	23.8%	41.9%	58.3%	31.9%	9.2%	17.3%	3.2%	14.8%	13.9%	4.9G	0.4M
Fast-SCNN [45]	24.8%	45.9%	60.0%	31.7%	9.7%	17.1%	6.0%	14.2%	13.8%	1.7G	1.1M
PSPNet18 [91]	28.2%	58.2%	66.8%	28.4%	13.0%	19.2%	6.2%	18.2%	15.6%	235.0G	17.5M
DRNet22 [82]	27.3%	53.0%	66.2%	19.7%	7.1%	16.3%	7.5%	20.5%	28.2%	136.5G	15.9M
ERFNet [48]	29.8%	64.7%	68.0%	25.9%	6.2%	22.0%	9.0%	19.2%	23.0%	30.3G	2.1M
CGNet [66]	25.8%	49.7%	60.1%	24.0%	9.9%	15.3%	4.6%	16.1%	26.5%	7.0G	0.5M
SwiftNet [40]	30.0%	55.7%	64.1%	29.2%	16.2%	22.9%	8.5%	21.1%	22.2%	41.7G	11.8M
SwaffNet [70]	35.4%	64.0%	68.5%	37.2%	10.7%	26.7%	13.1%	27.8%	35.6%	41.8G	11.9M
ERF-PSPNet [69]	34.0%	66.3%	70.5%	36.5%	6.4%	24.1%	9.4%	26.5%	32.0%	26.6G	2.5M
PASS (ERF-PSPNet) [69]	64.7%	87.3%	80.0%	61.4%	71.1%	49.9%	72.2%	37.5%	57.9%	91.6G	2.5M
OOSS (ERF-PSPNet) [71]	56.1%	87.2%	79.3%	60.8%	28.0%	38.1%	54.5%	48.8%	52.2%	26.6G	2.5M
Our omni-supervised (ERF-PSPNet)	66.8%	90.5%	82.7%	65.6%	70.5%	51.5%	58.2%	62.0%	53.1%	26.6G	2.5M
With attention (ECANet)	67.7%	90.4%	83.7%	68.4%	67.8%	52.1%	61.4%	54.5%	63.3%	27.8G	2.6M
With fusion (ECANet)	69.0%	90.6%	85.7%	68.0%	67.9%	52.1%	66.0%	59.3%	62.3%	27.8G	2.6M

Table 2. Accuracy analysis on WildPASS.

Network	Vistas-trained	Omni-supervised	Multi-space fused
ERF-PSPNet [69]	34.0%	66.8%	68.1%
Fast-SCNN [45]	24.8%	45.4%	45.8%
CGNet [66]	25.8%	48.2%	48.8%
DRNet22 [82]	27.3%	64.7%	66.0%
Our ECANet	39.7%	67.7%	69.0%

Table 3. Analysis of our approach for various efficient networks.

number of Parameters. It can be seen that they generally achieve unsatisfactory results. For example, DenseASPP is reported to have a mIoU of 65.8% on Vistas but degrades by more than 30.0% when taken to our challenging WildPASS. State-of-the-art panoramic segmentation frameworks PASS [69] and OOSS [71] improve upon the ERF-PSPNet baseline. However, PASS needs to separate the panorama into multiple partitions and thereby requires nearly 4 times of MACs, while OOSS is limited without materializing diverse panoramas in data distillation. Our proposed multi-source omni-supervised approach outperforms both of them while maintaining the high efficiency of ERF-PSPNet. With concurrent attention, the proposed ECANet surpasses ERF-PSPNet and our multi-space fusion elevates mIoU to 69.0%, outperforming all previous methods.

Effectiveness for different networks. Furthermore, we experiment with various efficient ConvNets to study how well our approach generalizes. As shown in Table 3, the multi-source omni-supervision is advantageous for all efficient networks, as the mIoU scores on WildPASS are improved by significant amounts, even up to 37.4%. The multi-space fusion is also consistently effective and the benefit is most pronounced for more accurate networks. These results demonstrate that our novel learning regimen is not strictly tied to a specific architecture and enables reliable predictions for omnidirectional imagery across the board.

4.5. Ablation Study

Ablation of multi-source omni-supervision. We first analyze our multi-source omni-supervision training method

Method	on Vistas	on WildPASS
Vistas-trained	61.6%	34.0%
Multi-source trained	63.0%	39.8%
Single-source fine-tuned	20.3%	41.1%
Multi-source fine-tuned	45.2%	60.7%
Omni-supervised (Pittsburgh) [71]	62.9%	56.1%
Omni-supervised (World-wide panoramas)	62.9%	66.8%
CLAN [35]	39.9%	38.3%
USSS [27]	49.5%	38.7%
PASS [69]	-	64.7%
Seamless [44]	-	34.0%
Our method (ERF-PSPNet)	62.9%	68.1%
Our method with attention (ECANet)	63.5%	69.0%

Table 4. Analysis of our multi-source learning method.

with ERF-PSPNet [69] on Vistas validation set and our WildPASS. As shown in Table 4, while the single-source Vistas-learned model is accurate in the training domain (61.6%), it only yields a mIoU of 34.0% on WildPASS. The IDD20K-trained model does not contain all the classes and thus no fair evaluation is possible, yet, the multi-source joint-training boosts the score to 39.8%. Fine-tuning on panoramas with generated annotations in both a single- or multi-source fashion improves the results on WildPASS. Omni-supervised learning by using the Pittsburgh dataset [71] is limited by less diverse panoramic images. In comparison, omni-supervision on world-wide panoramas largely improves the mIoU to 66.8%, demonstrating the importance of diverse panoramic images in the data distillation process. We further compare to the state-of-the-art domain adaptation method CLAN [35] by adapting from pinhole to panoramic imagery and the multi-source semi-supervised method USSS [27]. Moreover, PASS [69] for panoramic- and Seamless [44] for panoptic segmentation enable a smooth holistic understanding. However, both of the above frameworks remain sub-optimal in contrast to our method with ERF-PSPNet (68.1%) and ECANet (69.0%) which reaches high performances via multi-source omni-supervision and multi-space fusion.

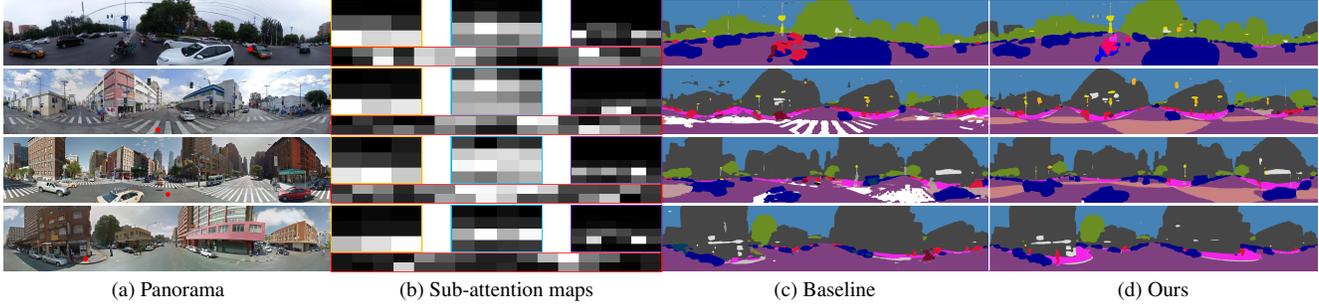


Figure 6. Qualitative examples of omnidirectional semantic segmentation: (a) Input panorama. (b) Sub-attention maps for the red point in (a), generated with our omni-supervised ECANet including the 2×16 map from the HSA- and the $\{3 \times 3, 4 \times 4, 6 \times 6\}$ maps from the PSA module. The brighter, the higher correlation to the region (maps normalized for visualization). For instance, the red point on the left sidewalk in the last row is correlated to the right sidewalk region according to the 2×16 map. (c) Baseline (ERF-PSPNet [69]). (d) Ours.

Network	mIoU	Memory	MACs	PARAMs
ERF-PSPNet [69]	34.0%	4.1M	4.6G	17K
ERFNet [48]	29.8%	224.0M	5.0G	189K
Non-local [64]	36.6%	261.3M	2.6G	38K
Dual attention [17]	36.7%	265.3M	2.9G	75K
Criss-cross attention [25]	38.3%	14.5M	0.7G	79K
Axial attention [61]	39.0%	22.5M	0.9G	104K
Our ECANet (HSA)	39.4%	5.3M	0.3G	43K
Our ECANet (PSA)	37.5%	14.7M	0.9G	116K
ECANet (HSA+PSA)	39.7%	20.0M	1.3G	159K

Table 5. Ablation of efficient concurrent attention. Computation metrics correspond to the context aggregation of single-source trained networks after ERFNet backbone (feature res. at 128×64).

Ablation of attention and fusion. Table 5 shows the mIoU, memory and computation requirement of our ECANet compared to its dissected versions using only HSA or PSA as well as non-local techniques from literature. It can be seen that our ECANet, while strengthening the horizontal context aggregation, is much more memory-efficient as compared to the non-local and dual attention modules. Moreover, ECANet clearly outperforms them as well as efficient axial- and criss-cross attention in omnidirectional segmentation. This demonstrates ECANet, by capturing essential omni-range contextual priors, is inherently robust in panoramic image segmentation. In the multi-source omni-supervised setting, we further compare different attention and context aggregation modules (Table 6). We experiment with $N = \{2, 4, 8\}$ segments and verify that a 4-segment ECANet is most accurate. Moreover, our multi-space fusion method exceeds calibration-bypassed fusion methods that simply take the class with maximum probability from both spaces or from the space with minimum variance.

4.6. Qualitative Analysis

Fig. 6 displays qualitative omnidirectional segmentation examples. For each panorama, we select a point and show their corresponding attention maps yielded from our ECANet. It can be observed that ECANet achieves more reliable segmentation than the pinhole-trained baseline, *e.g.*, ours enables complete segmentation of the crosswalk rather

Network/Method	mIoU	Network/Module	mIoU
ERF-PSPNet [69]	66.8%	Non-local [64]	62.6%
ECANet (HSA, 2 Segs)	66.8%	Dual attention [17]	65.0%
ECANet (HSA, 4 Segs)	67.4%	Criss-cross attention [25]	66.3%
ECANet (HSA, 8 Segs)	66.7%	Axial attention [61]	64.3%
ECANet (PSA)	67.2%	Height-driven attention [8]	64.8%
ECANet (HSA+PSA)	67.7%	Point-wise attention [92]	65.4%
Fusion (Max Prob.)	68.8%	ASPP [4]	65.7%
Fusion (Min Var.)	68.4%	BFP [14]	65.6%
Our Fusion	69.0%	Strip pooling [19]	63.7%

Table 6. Accuracy analysis of efficient concurrent attention and fusion methods in the multi-source omni-supervised setting, and comparison against state-of-the-art context aggregation modules.

than wrongly classifies into general road markings. The 3×3 attention map (orange) learned to highlight regions within the same height, which shows that it is meaningful to emphasize horizontal dependencies, forming a more discriminative feature representation. The 4×4 (blue) and 6×6 (purple) maps handle more detailed dependencies. On the other hand, the 2×16 (red) attention map captures semantically related regions in the horizontal direction stretching across 360° . For example, for the marked points in the first and last row, the spotlighted regions correspond to the strongly correlated homogeneous semantics, *i.e.*, other cars and sidewalks, whose features are passed back, significantly improving the consistency of the segmentation.

5. Conclusion

We look into omnidirectional image segmentation from the context-aware perspective. To bridge the gap in terms of FoV and structural distribution, our proposed ECANet captures inherent omni-range dependencies that stretch across 360° , whose generalization is optimized via multi-source omni-supervision and multi-space fusion. To foster progress in panoramic semantic perception, we establish and extensively evaluate models on WildPASS, a benchmark that captures diverse scenes from all around the world. Comprehensive experimental results show the proposed methods significantly elevate state-of-the-art accuracy for high-efficiency ConvNets on PASS and WildPASS.

References

- [1] 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*, 2017. 1, 2, 6, 7
- [2] Ignas Budvytis, Marvin Teichmann, Tomas Vojir, and Roberto Cipolla. Large scale joint semantic re-localisation and scene understanding via globally unique instance coordinate regression. In *BMVC*, 2019. 3
- [3] Abhishek Chaurasia and Eugenio Culurciello. Linknet: Exploiting encoder representations for efficient semantic segmentation. In *VCIP*, 2017. 6
- [4] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 2, 8
- [5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 2
- [6] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. A²-nets: Double attention networks. In *NeurIPS*, 2018. 2
- [7] Yuhao Chen, Ming Yang, Chunxiang Wang, and Bing Wang. 3d semantic modelling with label correction for extensive outdoor scene. In *IV*, 2019. 1
- [8] Sungha Choi, Joanne Taery Kim, and Jaegul Choo. Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks. In *CVPR*, 2020. 1, 2, 8
- [9] 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 *CVPR*, 2016. 4, 5
- [10] 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° panoramic imagery. In *ECCV*, 2018. 6
- [11] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, Bing Hu, and Chunxiang Wang. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 3
- [12] Liuyuan Deng, Ming Yang, Ye-qiang Qian, Chunxiang Wang, and Bing Wang. CNN based semantic segmentation for urban traffic scenes using fisheye camera. In *IV*, 2017. 3
- [13] Shuai Di, Qi Feng, Chun-Guang Li, Mei Zhang, Honggang Zhang, Semir Elezovikj, Chiu C. Tan, and Haibin Ling. Rainy night scene understanding with near scene semantic adaptation. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 6
- [14] Henghui Ding, Xudong Jiang, Ai Qun Liu, Nadia Magnenat-Thalmann, and Gang Wang. Boundary-aware feature propagation for scene segmentation. In *ICCV*, 2019. 8
- [15] Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. 1
- [16] Jun Fu, Jing Liu, Jie Jiang, Yong Li, Yongjun Bao, and Hanqing Lu. Scene segmentation with dual relation-aware attention network. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 2
- [17] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *CVPR*, 2019. 2, 3, 6, 7, 8
- [18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 2
- [19] Qibin Hou, Li Zhang, Ming-Ming Cheng, and Jiashi Feng. Strip pooling: Rethinking spatial pooling for scene parsing. In *CVPR*, 2020. 2, 3, 8
- [20] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *CVPR*, 2018. 2
- [21] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. AC-NET: Attention based network to exploit complementary features for RGBD semantic segmentation. In *ICIP*, 2019. 2
- [22] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 2
- [23] Lang Huang, Yuhui Yuan, Jianyuan Guo, Chao Zhang, Xilin Chen, and Jingdong Wang. Interlaced sparse self-attention for semantic segmentation. *arXiv:1907.12273*, 2019. 2
- [24] Shaofei Huang, Si Liu, Tianrui Hui, Jizhong Han, Bo Li, Jiashi Feng, and Shuicheng Yan. Ordnet: Capturing omni-range dependencies for scene parsing. *IEEE Transactions on Image Processing*, 2020. 2
- [25] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019. 2, 3, 8
- [26] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. *arXiv:2103.00868*, 2021. 1
- [27] Tarun Kalluri, Girish Varma, Manmohan Chandraker, and C. V. Jawahar. Universal semi-supervised semantic segmentation. In *ICCV*, 2019. 7
- [28] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6
- [29] Hanchao Li, Pengfei Xiong, Jie An, and Lingxue Wang. Pyramid attention network for semantic segmentation. In *BMVC*, 2018. 2
- [30] Xia Li, Zhisheng Zhong, Jianlong Wu, Yibo Yang, Zhouchen Lin, and Hong Liu. Expectation-maximization attention networks for semantic segmentation. In *ICCV*, 2019. 2
- [31] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 2
- [32] Yazhou Liu, Yuliang Chen, Pongsak Lasang, and Quansen Sun. Covariance attention for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2

- [33] Shao-Yuan Lo, Hsueh-Ming Hang, Sheng-Wei Chan, and Jing-Jih Lin. Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In *MM Asia*, 2019. 6
- [34] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1, 2
- [35] Yawei Luo, Liang Zheng, Tao Guan, Junqing Yu, and Yi Yang. Taking a closer look at domain shift: Category-level adversaries for semantics consistent domain adaptation. In *CVPR*, 2019. 7
- [36] Sachin Mehta, Mohammad Rastegari, Anat Caspi, Linda G. Shapiro, and Hannaneh Hajishirzi. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In *ECCV*, 2018. 6
- [37] Kenichi Narioka, Hiroki Nishimura, Takayuki Itamochi, and Teppei Inomata. Understanding 3d semantic structure around the vehicle with monocular cameras. In *IV*, 2018. 3
- [38] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kotschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *ICCV*, 2017. 2, 3, 4, 6
- [39] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *ICCV*, 2015. 2
- [40] Marin Oršić, Ivan Krešo, Petra Bevanđić, and Siniša Šegvić. In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In *CVPR*, 2019. 1, 6, 7
- [41] Bowen Pan, Jiankai Sun, Ho Yin Tiga Leung, Alex Andonian, and Bolei Zhou. Cross-view semantic segmentation for sensing surroundings. *IEEE Robotics and Automation Letters*, 2020. 3
- [42] Adam Paszke, Abhishek Chaurasia, Sangpil Kim, and Eugenio Culurciello. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv:1606.02147*, 2016. 6, 7
- [43] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters - improve semantic segmentation by global convolutional network. In *CVPR*, 2017. 2
- [44] Lorenzo Porzi, Samuel Rota Bulò, Aleksander Colovic, and Peter Kotschieder. Seamless scene segmentation. In *CVPR*, 2019. 7
- [45] Rudra P. K. Poudel, Stephan Liwicki, and Roberto Cipolla. Fast-scnn: Fast semantic segmentation network. In *BMVC*, 2019. 3, 6, 7
- [46] Ilija Radosavovic, Piotr Dollár, Ross B. Girshick, Georgia Gkioxari, and Kaiming He. Data distillation: Towards omniscient supervised learning. In *CVPR*, 2018. 2, 3, 4
- [47] Thomas Roddick and Roberto Cipolla. Predicting semantic map representations from images using pyramid occupancy networks. In *CVPR*, 2020. 3
- [48] Eduardo Romera, Jose M. Alvarez, Luis Miguel Bergasa, and Roberto Arroyo. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2018. 3, 6, 7, 8
- [49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2
- [50] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael S. Bernstein, Alexander C. Berg, and Fei-Fei Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 2015. 6
- [51] Haiwei Sang, Qiuha Zhou, and Yong Zhao. Pcanet: Pyramid convolutional attention network for semantic segmentation. *Image and Vision Computing*, 2020. 2
- [52] Ahmed Rida Sekkat, Yohan Dupuis, Pascal Vasseur, and Paul Honeine. The omniscene dataset. In *ICRA*, 2020. 2
- [53] Dingguo Shen, Yuanfeng Ji, Ping Li, Yi Wang, and Di Lin. Ranet: Region attention network for semantic segmentation. In *NeurIPS*, 2020. 2
- [54] Shuran Song, Andy Zeng, Angel X. Chang, Manolis Savva, Silvio Savarese, and Thomas A. Funkhouser. Im2pano3d: Extrapolating 360° structure and semantics beyond the field of view. In *CVPR*, 2018. 3
- [55] Lei Sun, Kailun Yang, Xinxin Hu, Weijian Hu, and Kaiwei Wang. Real-time fusion network for RGB-D semantic segmentation incorporating unexpected obstacle detection for road-driving images. *IEEE Robotics and Automation Letters*, 2020. 2
- [56] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV*, 2018. 3
- [57] Michael Trembl, José Arjona-Medina, Thomas Unterthiner, Rupesh Durgesh, Felix Friedmann, Peter Schuberth, Andreas Mayr, Martin Heusel, Markus Hofmarcher, Michael Widrich, Ulrich Bodenhofer, Bernhard Nessler, and Sepp Hochreiter. Speeding up semantic segmentation for autonomous driving. In *NeurIPS Workshop*, 2016. 6
- [58] Girish Varma, Anbumani Subramanian, Anoop M. Nambodiri, Manmohan Chandraker, and C. V. Jawahar. IDD: A dataset for exploring problems of autonomous navigation in unconstrained environments. In *WACV*, 2019. 2, 3, 4, 6
- [59] Arun Balajee Vasudevan, Dengxin Dai, and Luc Van Gool. Semantic object prediction and spatial sound super-resolution with binaural sounds. In *ECCV*, 2020. 1
- [60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 4
- [61] Huiyu Wang, Yukun Zhu, Bradley Green, Hartwig Adam, Alan L. Yuille, and Liang-Chieh Chen. Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In *ECCV*, 2020. 2, 8
- [62] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *CVPR*, 2020. 2
- [63] Shenlong Wang, Min Bai, Gellért Mátyus, Hang Chu, Wenjie Luo, Bin Yang, Justin Liang, Joel Cheverie, Sanja Fidler, and Raquel Urtasun. Torontocity: Seeing the world with a million eyes. In *ICCV*, 2017. 2

- [64] Xiaolong Wang, Ross B. Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *CVPR*, 2018. 2, 3, 4, 8
- [65] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018. 2
- [66] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Yongdong Zhang. Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Transactions on Image Processing*, 2021. 3, 6, 7
- [67] Kaite Xiang, Kailun Yang, and Kaiwei Wang. Polarization-driven semantic segmentation via efficient attention-bridged fusion. *Optics Express*, 2021. 2
- [68] Yuanyou Xu, Kaiwei Wang, Kailun Yang, Dongming Sun, and Jia Fu. Semantic segmentation of panoramic images using a synthetic dataset. In *SPIE*, 2019. 3
- [69] Kailun Yang, Xinxin Hu, Luis Miguel Bergasa, Eduardo Romera, and Kaiwei Wang. PASS: Panoramic annular semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 2, 3, 4, 5, 6, 7, 8
- [70] Kailun Yang, Xinxin Hu, Hao Chen, Kaite Xiang, Kaiwei Wang, and Rainer Stiefelwagen. DS-PASS: Detail-sensitive panoramic annular semantic segmentation through swafnet for surrounding sensing. In *IV*, 2020. 1, 3, 6, 7
- [71] Kailun Yang, Xinxin Hu, Yicheng Fang, Kaiwei Wang, and Rainer Stiefelwagen. Omnisupervised omnidirectional semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 2020. 1, 3, 4, 5, 6, 7
- [72] Kailun Yang, Xinxin Hu, and Rainer Stiefelwagen. Is context-aware CNN ready for the surroundings? panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, 2021. 2
- [73] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *CVPR*, 2018. 1, 2, 6, 7
- [74] Yaozu Ye, Kailun Yang, Kaite Xiang, Juan Wang, and Kaiwei Wang. Universal semantic segmentation for fisheye urban driving images. In *SMC*, 2020. 3
- [75] Minghao Yin, Zhuliang Yao, Yue Cao, Xiu Li, Zheng Zhang, Stephen Lin, and Han Hu. Disentangled non-local neural networks. In *ECCV*, 2020. 2
- [76] Senthil Kumar Yogamani, Christian Witt, Hazem Rashed, Sanjaya Nayak, Saquib Mansoor, Padraig Varley, Xavier Perrotton, Derek O’Dea, Patrick Pérez, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Sumanth Chennupati, Michal Uricár, Stefan Milz, Martin Simon, and Karl Amende. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *ICCV*, 2019. 2
- [77] Changqian Yu, Yifan Liu, Changxin Gao, Chunhua Shen, and Nong Sang. Representative graph neural network. In *ECCV*, 2020. 2
- [78] Changqian Yu, Jingbo Wang, Changxin Gao, Gang Yu, Chunhua Shen, and Nong Sang. Context prior for scene segmentation. In *CVPR*, 2020. 2
- [79] Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *ECCV*, 2018. 3, 6
- [80] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *CVPR*, 2020. 5
- [81] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. In *ICLR*, 2015. 2
- [82] Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser. Dilated residual networks. In *CVPR*, 2017. 6, 7
- [83] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *arXiv:1809.00916*, 2018. 2
- [84] Oliver Zendel, Katrin Honauer, Markus Murschitz, Daniel Steinger, and Gustavo Fernández Domínguez. Wilddash - creating hazard-aware benchmarks. In *ECCV*, 2018. 6
- [85] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *ICCV*, 2019. 3
- [86] Hang Zhang, Kristin J. Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Amrith Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 2
- [87] Hang Zhang, Han Zhang, Chenguang Wang, and Junyuan Xie. Co-occurrent features in semantic segmentation. In *CVPR*, 2019. 2
- [88] Jiaming Zhang, Kailun Yang, and Rainer Stiefelwagen. IS-SAFE: Improving semantic segmentation in accidents by fusing event-based data. *arXiv:2008.08974*, 2020. 2, 6
- [89] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip H. S. Torr. Dual graph convolutional network for semantic segmentation. In *BMVC*, 2019. 2
- [90] Hengshuang Zhao, Xiaojuan Qi, Xiaoyong Shen, Jianping Shi, and Jiaya Jia. Icnets for real-time semantic segmentation on high-resolution images. In *ECCV*, 2018. 6
- [91] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 2, 6, 7
- [92] Hengshuang Zhao, Yi Zhang, Shu Liu, Jianping Shi, Chen Change Loy, Dahua Lin, and Jiaya Jia. Psanet: Pointwise spatial attention network for scene parsing. In *ECCV*, 2018. 2, 8
- [93] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 3
- [94] Zhen Zhu, Mengdu Xu, Song Bai, Tengpeng Huang, and Xiang Bai. Asymmetric non-local neural networks for semantic segmentation. In *ICCV*, 2019. 2