Capturing Omni-Range Context for Omnidirectional Segmentation

Kailun Yang1 Jiaming Zhang1 Simon Reiß1,2 Xinxin Hu3 Rainer Stiefelhagen1

1CV:HCI, Karlsruhe Institute of Technology 2ZEISS 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 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
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 Semantic 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 groundbreaking work 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 transfer [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 \(O((H_e \times W_e) \times (H_e \times W_e)) \sim O(H_e^2W_e^2)\). \(H_e \times W_e\) 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 $I \in \mathbb{R}^{C_i \times H_x \times W_x}$ 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 $S \in \mathbb{R}^{C \times H \times W}$ has a smaller size with $(H, W) = (\frac{H}{N}, W_x)$, which is fed into convolution layers to generate feature maps of query $Q$, key $K$ and value $V$ to carry out self-attention [60]. This reinforces the horizontal context aggregation and reduces the complexity to $O(N \times (\frac{H}{N} \times W_x) \times \frac{H}{N} \times W_x) \sim O(H^2 W_x^2 / N)$.

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 $K \in \mathbb{R}^{C \times H \times W}$ and value feature $V \in \mathbb{R}^{C \times H \times W}$. 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 $A$, is then computed from the reshaped query feature $Q' \in \mathbb{R}^{(H \times W) \times C}$ and the new key feature $K' \in \mathbb{R}^{C \times (\hat{H} \times \hat{W})}$, $\hat{Q}' \times K' \rightarrow A \in \mathbb{R}^{(H \times W) \times (\hat{H} \times \hat{W})}$

In this manner, the association between each pixel and striped-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 $V' \in \mathbb{R}^{C_v \times (\hat{H} \times \hat{W})}$ as follows:

$$V' \times A^T \rightarrow S' \in \mathbb{R}^{C_v \times (H \times W)}$$

Finally, all segments $S' \in \mathbb{R}^{C_v \times H \times W}$ are concatenated along the vertical dimension $H$ amounting to the horizontally-output driven output $O \in \mathbb{R}^{C_v \times H \times W_e}$. With this architectural design, we are able to greatly reduce the complexity to $O(N \times (\frac{H}{N} \times W_x) \times (\hat{H} \times \hat{W})) \sim O(H \times W_x \times \hat{H} \times \hat{W})$, as $\hat{H} \hat{W} \ll H \times W_e$. 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 $\sum_i O(H \times W \times \hat{H}_i \times \hat{W}_i)$, a considerable 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 assembled, producing the annotation. While time-consuming due to the complex PASS
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 Vista-space) is used for the inferred omnidirectional perception. We assess this to be sub-optimal as it disregards the overlap in labels of the multi-source 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 \( C_j \): \( C_{IN} = \cap_{j=1}^{6} C_j \).

Based on the unnormalized output \( z_j^i \) for the \( i \)th pixel as produced by the trained \( j \)th output-head, the predicted class \( y_j^i \) is known as \( y_j^i = \arg \max_{c \in C_j} z_j^i,c \). However, considering other candidate heads, the prediction will be updated as:

\[
y_j^i = \arg \max_{c \in C_{IN}} z_j^i,c^* \tag{3}
\]

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 \( j_{\text{unc}}^* = \min_{j \in J} \left( \text{var}(z_j^i) \right) \), where uncertainty is modeled by the variance \( \text{var}(z_j^i) = (1/C_j) \sum_{c=1}^{C_j} (z_j^i,c - \bar{z}_j^i)^2 \). 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: \( j_{\text{conf}}^* = \max_{j \in J} \max \sigma(z_j^i,c) \) 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: \( j_{\text{cal}}^* = \max_{j \in J} \frac{\max \sigma(z_j^i,c)}{\max_2 \sigma(z_j^i,c)} \). Compared to using all values in \( j_{\text{conf}}^* \) or just the maximum one in \( j_{\text{conf}}^* \), 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,
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 $\times$ 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 $(\tilde{H}, \tilde{W}) = (2, 16)$ in HSA and spatial pooling of $(\tilde{H}, \tilde{W}) = \{(3, 3), (4, 4), (6, 6)\}$ in PSA. ECANet is trained under Adam optimization [28] with a Weight Decay of $2 \times 10^{-4}$ and a Learning Rate of $1 \times 10^{-4}$ that decreases exponentially over 200 epochs. We train with a batch-size of 12 and a resolution of 1024 $\times$ 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 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-paddling [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
We first analyze our multi-source omni-supervision training method with ERF-PSPNet [69] on Vistas validation set and our multi-source learning method. 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-supervised learning 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 panoramic 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.

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

| Network            | Vistas-trained | Omni-supervised | Multi-space fused |
|--------------------|----------------|------------------|-------------------|
| ERF-PSPNet [69]    | 61.6%          | 63.0%            | 60.7%             |
| Fast-SCNN [45]     | 62.9%          | 64.2%            | 61.9%             |
| CGNet [66]         | 58.4%          | 59.7%            | 57.0%             |
| DRNet22 [82]       | 67.7%          | 69.0%            | 66.3%             |
| Our ECANet         | 69.0%          | 70.5%            | 68.0%             |

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

| Method               | [on Vistas] | [on WildPASS] |
|----------------------|-------------|---------------|
| Vistas-trained       | 61.6%       | 63.0%         |
| Multi-source trained | 60.7%       | 65.8%         |
| Single-source fine-tuned | 52.1% | 53.1%         |
| Multi-source fine-tuned | 45.2% | 50.9%         |

Conclusion

Our multi-source training framework consistently demonstrates superior performance across a variety of network architectures and datasets. The multi-source omni-supervision and multi-space fusion components are crucial for improving segmentation accuracy and robustness. The effectiveness is particularly pronounced with more accurate networks, and the method is demonstrated to be scalable and effective for diverse panoramic domains.
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 $2\times16$ map from the HSA- and the $\{3\times3, 4\times4, 6\times6\}$ 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 $2\times16$ map. (c) Baseline (ERF-PSPNet [69]). (d) Ours.

| Network/Method                     | mIoU   | Network/Module | mIoU   |
|------------------------------------|--------|----------------|--------|
| ERF-PSPNet [69]                    | 68.8%  | Non-local [64] | 62.6%  |
| ERFNet [48]                        | 66.8%  | Dual attention [17] | 65.0%  |
| Non-local [64]                     | 67.4%  | Cross-cross attention [25] | 66.3%  |
| Dual attention [17]                | 67.7%  | Axial attention [61] | 64.3%  |
| Criss-cross attention [25]         | 67.2%  | Height-driven attention [8] | 64.8%  |
| Axial attention [61]               | 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.

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 than wrongly classifies into general road markings. The $3\times3$ 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 $4\times4$ (blue) and $6\times6$ (purple) maps handle more detailed dependencies. On the other hand, the $2\times16$ (red) attention map captures semantically related regions in the horizontal direction stretching across $360^\circ$. 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 $360^\circ$, 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

[5] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous convolution for semantic image segmentation. In *ECCV*, 2018. 2

[6] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, 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, Yeqiang 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] Xin 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] Kaiying He, Xiangyu Zhang, Shaqing 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. ACNET: 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, Xin Lin, and Jingdong Wang. Interfaced 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. Ornet: Capturing omni-range dependencies for scene parsing. *IEEE Transactions on Image Processing*, 2020. 2

[25] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunhao Wei, and Wen Yu Liu. Cnnet: 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] Diedrik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[29] Hanxiao 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 Zhang. Covariance attention for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 2
