

FishDreamer: Towards Fisheye Semantic Completion via Unified Image Outpainting and Segmentation

Hao Shi^{1,5,*}, Yu Li^{2,*}, Kailun Yang^{3,†}, Jiaming Zhang^{2,4}, Kunyu Peng², Alina Roitberg², Yaozu Ye¹, Huajian Ni⁵, Kaiwei Wang^{1,†}, Rainer Stiefelhagen²

¹Zhejiang University ²Karlsruhe Institute of Technology ³Hunan University ⁴University of Oxford ⁵Supremind

Abstract

This paper raises the new task of **Fisheye Semantic Completion (FSC)**, where dense texture, structure, and semantics of a fisheye image are inferred even beyond the sensor field-of-view (FoV). Fisheye cameras have larger FoV than ordinary pinhole cameras, yet its unique special imaging model naturally leads to a blind area at the edge of the image plane. This is suboptimal for safety-critical applications since important perception tasks, such as semantic segmentation, become very challenging within the blind zone. Previous works considered the out-FoV outpainting and in-FoV segmentation separately. However, we observe that these two tasks are actually closely coupled. To jointly estimate the tightly intertwined complete fisheye image and scene semantics, we introduce the new **FishDreamer** which relies on successful ViTs enhanced with a novel Polar-aware Cross Attention module (PCA) to leverage dense context and guide semantically-consistent content generation while considering different polar distributions. In addition to the contribution of the novel task and architecture, we also derive *Cityscapes-BF* and *KITTI360-BF* datasets to facilitate training and evaluation of this new track. Our experiments demonstrate that the proposed **FishDreamer** outperforms methods solving each task in isolation and surpasses alternative approaches on the **Fisheye Semantic Completion**. Code and datasets are publicly available at [FishDreamer](#).

1. Introduction

Benefiting from a larger field-of-view (FoV), fisheye cameras have been widely used in autonomous driving and mobile robots [13, 22, 43, 48, 72]. However, due to the special optical design of the fisheye camera, there are invalid black areas at the edge of the image plane. Interestingly, humans have a natural ability to infer complete semantic

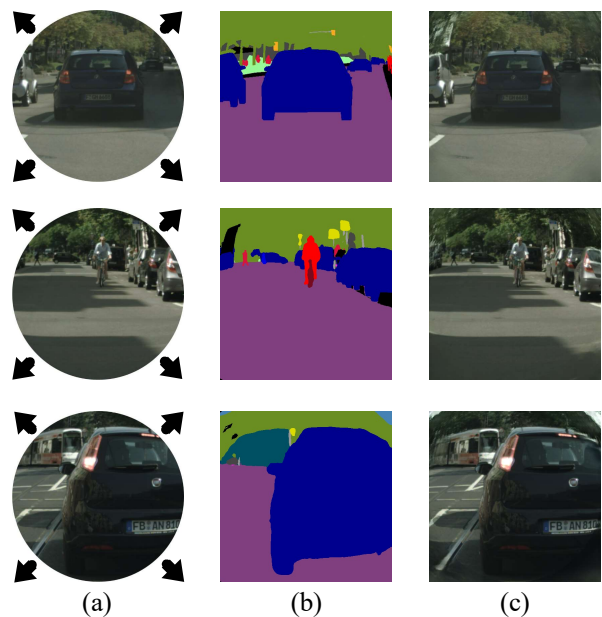


Figure 1. Illustrations of our proposed **Fisheye Semantic Completion**. (a) Input single-view fisheye image. (b)(c) Fisheye semantic completion result: our model jointly predicts texture, structure and object categories for each pixel beyond the fisheye vision.

information from partial visual observations [31, 41] (e.g., a partially occluded car) to navigate and interact in the real world. Similarly, for an ego-view agent, the ability to estimate the full field-of-view of a given scene is beneficial for mid-level tasks such as obstacle avoidance [20], while perceiving semantic concepts is a prerequisite for complex cognitive tasks such as high-level scene understanding, planning the next step or answering questions about the space [5, 15].

With this motivation, our goal is to build a model that can simultaneously complete the missing image areas and generate predictions for semantic object categories from a single-shot fisheye image in an end-to-end manner. We refer to this novel task as “Fisheye Semantic Completion”, with an overview of the proposed problem given in Fig. 1. Our key idea is grounded by the observation that the distri-

*The first two authors contribute equally to this work.

†Corresponding author (e-mail: kailun.yang@hnu.edu.cn, wangkaiwei@zju.edu.cn).

bution of pixel values of an entity within an image is tightly coupled to its semantic label. Therefore, the two problems of outpainting the content outside the fisheye camera’s FoV and the semantics of objects are strongly conjugated, which we believe is vital for achieving good performance in both tasks. In other words, if we know the semantic categories of an incomplete object, we can predict its pixel pattern on the image plane even without direct observation (*e.g.*, seeing a tree trunk appearing in the FoV and then inferring the presence of tree canopy outside). Conversely, having a complete observation of an object can help us recognize its semantic class more accurately.

To achieve this goal, we must overcome several key challenges: First, how to effectively take advantage of the strong coupling characteristics of these two sub-tasks (*i.e.* fisheye outpainting and segmentation) to realize a win-win situation? Second, since existing fisheye semantic segmentation datasets cannot provide images and semantic ground truth outside the FoV, how can we obtain fisheye beyond-FoV training data with complete annotations?

To address the first challenge, we propose **FishDreamer**, which benefits from the successful Visual Transformer (ViT) structure as the backbone, and integrates a novel Polar-aware Cross Attention module (PCA) to enhance the flow of visual cues between two sub-tasks. Specifically, PCA takes into account the unique polar distribution and distortion patterns of fisheye cameras, considers the heterogeneity of different polar coordinate locations when querying relevant visual features, and leverages the rich semantic context from the semantic head to guide the outpainting head in hallucinating semantically continuous and plausible content outside the fisheye FoV. As for the data challenge, we leverage the popular Cityscapes [7] and KITTI360 [28] semantic segmentation datasets via pinhole-to-fisheye projection and derive the new CityScapes-BF and KITTI360-BF as beyond-FoV benchmark variants, therefore enabling training and evaluation of fisheye semantic completion.

Extensive experiments demonstrate that the proposed FishDreamer, which jointly learns semantics and content outside the fisheye FoV, outperforms approaches that address the two sub-tasks in isolation. The proposed PCA module, which focuses on the natural polarity distribution of the fisheye image and extracts visual cues from semantic priors, significantly improves performance. On the derived CityScapes-BF benchmark, FishDreamer achieves state-of-the-art performance with a mIoU of 54.54% and a peak-SNR of 25.05dB, a 0.42dB performance gain from the best published result (24.63dB). FishDreamer also surpasses alternative approaches on KITTI360-BF as it hallucinates more realistic content in the blind area of the fisheye and gives clearer and sharper segmentation results.

In summary, we deliver the following contributions:

- We raise the new Fisheye Semantic Completion task,

which extends beyond fisheye vision and enables outpainting and semantic segmentation of the full scene.

- We establish the CityScapes-BF and KITTI360-BF benchmarks and validate existing models on this new fisheye semantic completion track.
- We propose FishDreamer, which utilizes a novel Polar-aware Cross Attention (PCA) module to effectively guide fisheye outpainting using semantic context.
- Extensive experiments demonstrate that the proposed FishDreamer outperforms alternative approaches that address the sub-tasks separately.

2. Related Work

Image outpainting. Image outpainting aims to generate the surrounding regions of the given visual content. Early parameter-free methods [49, 56, 76] are data-driven and are based on very large image databases or require input reference frames, which retrieve relevant image features to warp and fill in regions-of-interest. Sabini *et al.* [45] first present the learning-based image outpainting via Generative Adversarial Network (GAN) [14] to enable outpainting in the horizontal direction. Subsequently, Wang *et al.* [58] proposes generating semantically coherent structures and textures using a context prediction network and a carefully designed loss function. The framework of Teterwak *et al.* [53], leverages semantic information extracted from a pretrained deep network to modulate the discriminator’s behavior for image extension. Yao *et al.* [69] implemented a sequence-to-sequence outpainting approach that relies on a transformer-based backbone, where the outpainting proportion and the network structure are bound. FlowLens [50] introduces a temporal clip propagation mechanism to expand the FoV of the pinhole camera outwards and the spherical camera inwards, respectively. RecRecNet [27] rectangles rectified wide-angle via curriculum learning with increasing degree of freedom. The work most closely related to ours is presumably FisheyeEX [25], which leverages an outpainting method specifically for elimination of fisheye blind areas.

Different from prior works, we focus on *fisheye semantic completion*, that seeks to generate *semantically coherent* visual content beyond the fisheye FoV by jointly learning the scene semantics and pixel patterns and considering the fisheye polar distributions. To the best of our knowledge, this is the first work that concurrently addresses the challenges of fisheye semantic segmentation and scene completion.

Beyond-FoV semantic segmentation. Early fisheye semantic segmentation methods [10, 46, 47] generate synthetic fisheye datasets based on existing pinhole semantic segmentation datasets. These techniques employ focal length augmentation to enhance their adaptability to real-world scenarios. In [4, 70], the degree of freedom in generating synthetic fisheye images is enlarged, transforming rectilinear images to fisheye images in a more comprehen-

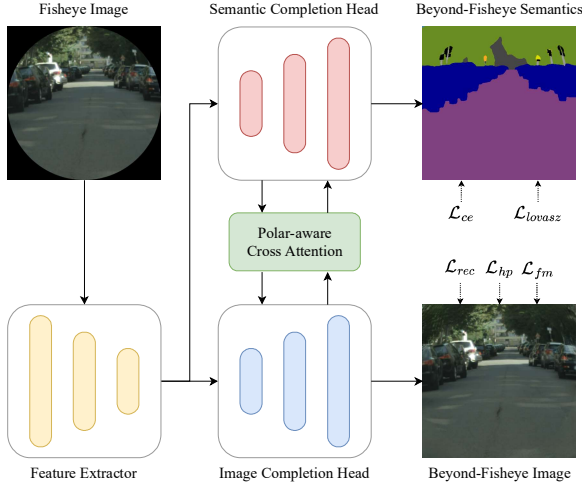


Figure 2. *FishDreamer: Fisheye semantic completion transformer.* Taking a single fisheye image as input, the model predicts pixel patterns and object labels beyond fisheye vision. Swin [33] is used as the backbone. The two sub-task heads interact each other via the proposed Polar-aware Cross Attention module bidirectionally.

sive way and enhancing the generalization on real fisheye images with various perspectives. In [1, 9, 16, 42], deformable components in CNNs are investigated to better adapt to wide-angle images. In [38, 61–64], the FoV is further expanded to 360° with panoramic or annular images. In [66, 67, 74, 75], wide-FoV-driven visual attention and distortion-aware transformer models are designed to learn long-range dependencies in panoramic images. In [21, 65, 78], knowledge distillation is studied on panoramic images. In [17, 21, 36, 51, 73], large-FoV semantic segmentation is revisited from a domain adaptation perspective by adapting from label-rich pinhole images to label-scare images such as fisheye images, panoramic images, and images reflected by convex mirrors. In [18, 37, 54], semantic segmentation is extended to panoptic segmentation on wide-FoV images with instance predictions. In [6, 18, 19], pixel-level contrastive learning is studied for wide-angle segmentation. In [2, 12, 23, 24], multi-task learning has been implemented on fisheye images such as object detection, depth estimation, and semantic segmentation. In contrast to these works, our work tackles fisheye semantic completion, which provides dense semantic information not only for the original wide-angle fisheye images but also beyond the field of view, giving rich semantic understanding of the expanded scene.

3. Methodology

3.1. Overview

In this section, we introduce FishDreamer – a novel approach capable of simultaneously achieving reliable extrapolation of fisheye images and a full-FoV semantic segmentation for both visible and previously unseen areas within fisheye images. As illustrated in Fig. 2, FishDreamer com-

prises four modules: the feature extraction backbone based on the Swin Transformer [33] (Sec. 3.2), the outpainting module (Sec. 3.3), the semantic decoder based on UPerNet [59] (Sec. 3.4) and the Polar-aware Cross Attention (PCA) mechanism (Sec. 3.5). Next, we will provide a comprehensive description of each module and their respective roles within the FishDreamer framework.

3.2. Feature Extractor

We begin by describing our feature extraction backbone, the Swin Transformer [33], which facilitates informative hierarchical feature learning and has been proven very effective, *e.g.*, in image and video classification [33, 34], activity recognition [39] and vanilla semantic segmentation [29]. Similar to other transformer-based models Swin leverages self-attention [55], but also employs a non-overlapping shifted window partitioning mechanism which enhances efficiency by focusing on the generated windows while preserving cross-window communication capabilities.

We utilize a four-stage Swin Transformer as our feature extraction backbone. In the first stage, multiple non-overlapping image patches are generated. As we address outpainting and semantic segmentation using 2D image data as input, a 2D shifted window pipeline is employed, operating within the 2D spatial domain. Assuming the 2D spatial dimensions of the input image are $[H, W]$ and the shifted window size is chosen as $[N, N]$, a total of $\frac{H}{N} \times \frac{W}{N}$ patches are extracted using the aforementioned window partitioning technique. Next, these patches are projected from \mathbb{R}^3 to \mathbb{R}^C using a linear projection layer. Next, we will describe further details of the Swin Transformer blocks.

The Swin Transformer block leverages its own Shifted-Window based Multi-head Self-Attention (SW-MSA) mechanism, as opposed to the standard Multi-head Self-Attention (MSA) found in ViT [11]. This approach mitigates the limitations of the vanilla ViT structure, specifically concerning the lack of cross-window connections and restricted model capacity. The workflow of the Swin Transformer block can be described as:

$$\begin{aligned} \hat{z}_{l-1} &= SW-MSA(LN(z_{l-1})), \\ z_l &= MLP(LN(\hat{z}_{l-1}) + \hat{z}_{l-1}), \end{aligned} \quad (1)$$

where SW-MSA stands for the Shifted-Window based Multi-head Self-Attention, LN indicates layer normalization, MLP is multi layer perception with GELU nonlinearity, l marks the layer number, \hat{z}_{l-1} is the attention output, z_{l-1} and z_l are the outputs of $\{l-1\}$ -th and $\{l\}$ -th modules, respectively. After each module there is a residual connection. All Swin blocks are equipped with the shifted window partitioning approach described above. Swin processes the image in a hierarchical manner, since it splits the input into non-overlapping patches and subsequently merges them at different resolutions.

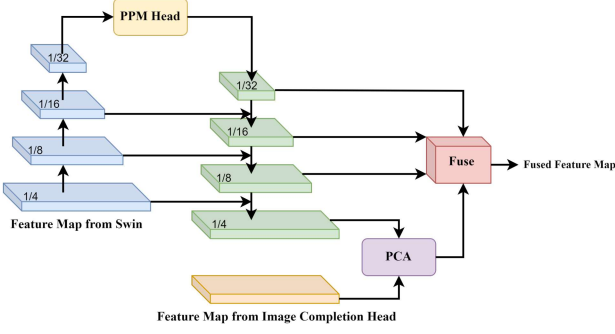


Figure 3. *UPerNet with PCA*. Hierarchical UPerNet with extra information from the image completion head.

3.3. Outpainting

Outpainting approaches typically employ several layers of deconvolution as the task head, and FishDreamer is no exception. Given that fisheye semantic completion seeks to accomplish both image completion and semantic completion simultaneously, we opt not to incorporate a more complex outpainting decoder in the design. Specifically, FishDreamer uses three layers of deconvolution, along with the PCA mechanism described in Sec. 3.5, to upsample and produce the feature map obtained from the feature extractor. This process generates the extrapolated output, which can be calculated as follows:

$$\begin{aligned} \hat{z}_s &= \text{ConvT}_i(z_s), \\ z &= \text{ConvT}_{i+1}(\hat{z}_s + \text{PCA}(\hat{z}_s, z_c)), \\ z &:= \text{ConvT}_{i+2}(z), \end{aligned} \quad (2)$$

where ConvT is the 2D transposed convolution operator with kernel size= 3×3 while the stride and padding are set as 2 and 1. i indicates the stage of the ConvT and is chosen as 0 in this work. PCA is the proposed Polar-aware Cross Attention module which will be detailed in Sec. 3.5, and z_s and z_c denote the feature extracted from the Swin Transformer-based backbone and the features from semantic completion head, \hat{z}_s denotes the feature map of the semantic segmentation branch after the first ConvT layer, z denotes the final merged feature map, respectively. With additional priors from the semantic completion head we can acquire a semantically coherent outpainting result.

3.4. Segmentation

Considering the hierarchical structure of the features obtained using the Swin Transformer backbone, more advanced techniques for dense prediction, such as Feature Pyramid Networks (FPN) [30] and U-Net [44], can be employed for linking the hierarchical information. In our work, we incorporate UPerNet [59] with our proposed PCA.

As depicted in Fig. 3, UPerNet features a top-down architecture with lateral connections that facilitate the fusion

of high-level semantic information with lower-level details. The model utilizes a Pyramid Pooling Module (PPM) [77] to achieve a larger receptive field and generate effective global prior representations. UPerNet has the capability to learn visual attributes for semantic segmentation and image completion at multiple levels. $\{F_1, F_2, F_3, F_4\}$ denotes the set of the resulting feature maps of each stage of the backbone. The corresponding downsampling rate for the $\{F_1, F_2, F_3, F_4\}$ are $\{4, 8, 16, 32\}$, respectively. In the decoder stage, the PPM is only used at the top of F_4 , where the resulting feature map can be denoted as \hat{F}_4 . Then, \hat{F}_4 is upsampled and sum with F_3 to progressively fuse multi-level features. The upsampling and downsampling rates are kept the same and finally four feature maps are obtained through the decoder stage. The corresponding feature maps for the four stages of the decoder are referred to as $\{\hat{F}_1, \hat{F}_2, \hat{F}_3, \hat{F}_4\}$. \hat{F}_4 (which is alternatively denoted as z_s) is then merged with the feature z_c from the image completion branch by using the PCA module explained in the next subsection. The final feature map can be obtained as follows:

$$z_{out} = \text{Fuse}(\hat{F}_1, \hat{F}_2, \hat{F}_3, z_{pca}), \quad (3)$$

where the z_{pca} denotes the output feature map of the PCA mechanism. We fuse the feature maps via concatenation.

3.5. Polar-aware Cross Attention

As image outpainting and semantic segmentation tasks are tightly intertwined, we design the novel Polar-aware Cross Attention (PCA) module to encourage the information flow between the two heads.

Given the two feature maps from the semantic segmentation head (z_s) and the image completion head (z_c), PCA initially constrains the polar distribution within each patch using a polar mask obtained from a newly designed polar-mask generator. This generator creates polar masks with varying quantities and radius. Let the total number of the generated masks be N_{mask} and the set of polar masks be $M = \{M_i \mid i \in [0, N_{mask}]\}$. A linear projection layer is employed for the masked feature map of each sub-task. The projected and masked patch partitions for both the feature maps from the semantic segmentation- (z_s) and the image completion heads (z_c) can be computed as follows:

$$z_s^*, z_c^* = L_s(M \odot z_s), L_c(M \odot z_c), \quad (4)$$

where the L_s and L_c indicate the linear projection layers of the semantic segmentation branch and the image completion branch, respectively. A multi-head cross-attention mechanism is utilized to integrate the focuses from the image completion branch into the semantic segmentation branch. To this intent, linear projection layers P_Q , P_K , and P_V are employed to compute the necessary query, key, and

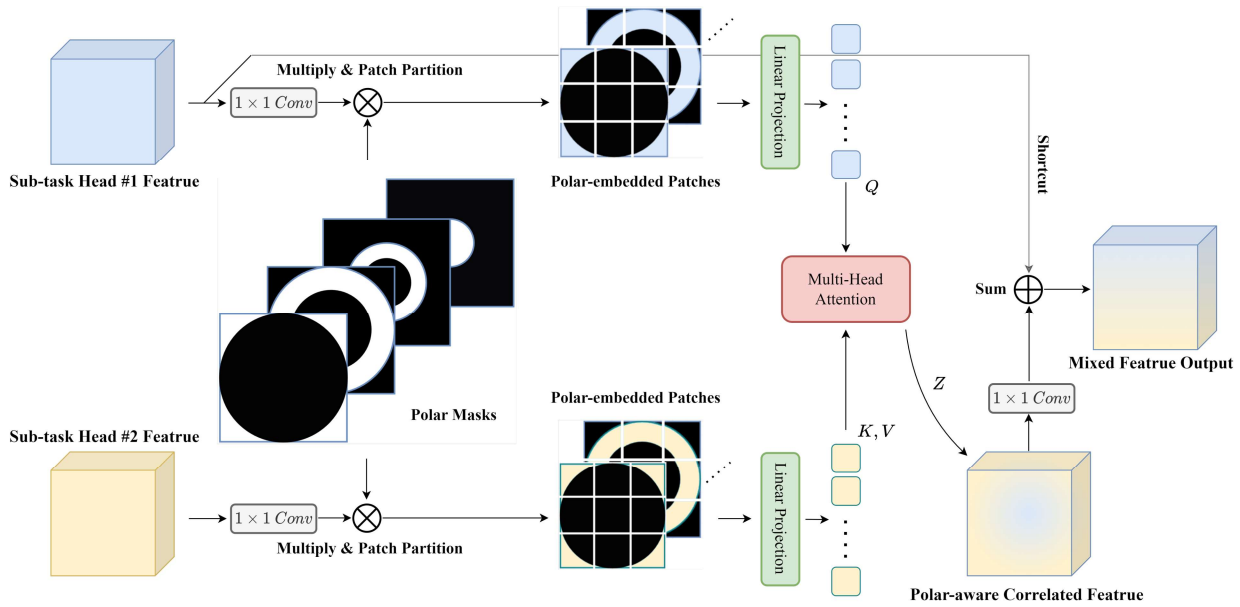


Figure 4. *Polar-aware Cross Attention module (PCA)*. Given the feature maps from two sub-task heads, PCA explicitly limits the polarity distribution in each patch by introducing N ring-shape masks, and exploit the multi-head attention to query polar-aware correlated feature from coupled sub-task feature map. Note this behavior can be either unidirectional or bidirectional.

value components for the MSA. The resulting merged feature map is then passed through an additional bottleneck layer to obtain the combined feature map. This final merged feature map is subsequently added to z_s^* to produce the ultimate output. This workflow can be formalized as:

$$z_{pca} = z_s^* + C_{BN}(SA(P_Q(z_s^*), P_K(z_c^*), P_V(z_c^*)), \quad (5)$$

where z_{pca} denotes the final mixed feature output of the proposed PCA mechanism, the C_{BN} denotes the bottleneck layer and SA denotes the self-attention mechanism which is calculated as $SA(Q, K, V) = \text{Softmax}(QK^T / \sqrt{\rho})V$ and ρ denotes the scale factor [11].

In comparison to the previous FishFormer work [68], which solely focuses on one task, *i.e.*, fisheye distortion correction, our model (shown in Fig. 4) simultaneously addresses two critical tasks for autonomous driving, namely semantic segmentation and image completion, which involves predicting the unseen regions of fisheye images by considering the entire scene. This blind zones information is crucial for autonomous vehicles in order to minimize potential risks, such as route planning and risk alerting within the blind zones of fisheye sensors. Our PCA mechanism is employed at the end of the model, merging feature maps from both the semantic segmentation and image completion heads. In contrast, the Layer Attention Mechanism (LAM) proposed by FishFormer, is integrated between every two transformer blocks, which is less efficient in terms of the number of attention blocks used. Additionally, while FishFormer encodes different annular slices, we force each to-

ken to encode features of a specific polar distribution. Tokens lacking valid features are discarded to enhance computational efficiency.

3.6. Training

Loss function. Multiple well-established loss functions are employed to simultaneously ensure the accuracy of blind-area prediction and complete-FoV semantic segmentation. Considering the full-scene image completion task of the fisheye image, we make use of a high receptive field perceptual loss, *i.e.*, \mathcal{L}_{hp} , an adversarial loss, *i.e.*, \mathcal{L}_{adv} , a reconstruction loss, *i.e.*, \mathcal{L}_{rec} and a feature matching loss, *i.e.*, \mathcal{L}_{fm} . Next, we discuss these loss functions in detail.

First, leverage the high receptive field perceptual loss [52] \mathcal{L}_{hp} , which calculates the difference between the feature maps of the predicted results and the target images. This loss function does not require an exact reconstruction workflow, which is particularly suitable for our case when addressing the challenge of limited information in the blind area. \mathcal{L}_{hp} can be calculated as:

$$\mathcal{L}_{hp}(z_p, z_t) = \mathcal{M}(D(\phi(z_p), \phi(z_t))^2), \quad (6)$$

where $D(\cdot)$ denotes an element-wise distance function (mean-squared error) loss and \mathcal{M} denotes the sequential two-stage mean operation. The \mathcal{L}_{hp} calculates the distance between the extracted features of the prediction (z_p) and the ground truth (z_t). ϕ denotes dilated convolutions. \mathcal{L}_{hp} does not require an exact reconstruction, which is a very good

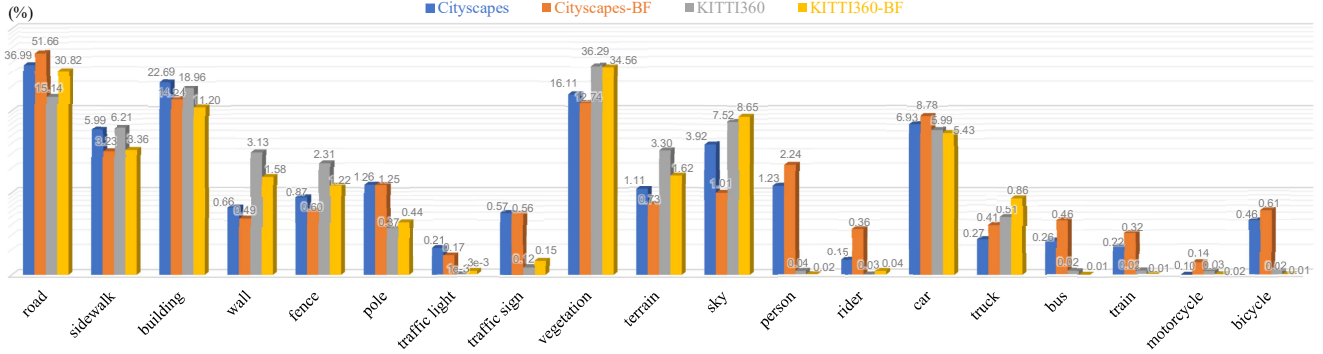


Figure 5. Distributions of Cityscapes [7], Cityscapes-BF, KITTI360 [28], and KITTI360-BF in terms of class-wise pixel percentage across the datasets. We use the logarithmic scaling of the vertical axis and insert the pixel frequency above the bar.

property in our case due to the lack of information of the blind area.

Second, an adversarial loss is used to ensure the preservation of local details. A discriminator, $D_\xi(\cdot)$, is used to distinguish between "real" and "fake" patches. The visible parts of the built images are marked as real, while patches that intersect with the blind area are marked as fake. We then compute the non-saturating adversarial loss \mathcal{L}_{adv} as:

$$\mathcal{L}_D = -\mathbb{E}_x[\log D_\xi(x)] - \mathbb{E}_{x,m}[\log D_\xi(\hat{x}) \odot M] - \mathbb{E}_{x,m}[\log(1 - D_\xi(\hat{x})) \odot (1 - M)], \quad (7)$$

$$\mathcal{L}_G = -\mathbb{E}_{x,m}[\log D_\xi(\hat{x})], \quad (8)$$

$$L_{adv} = \text{sg}_\theta(\mathcal{L}_D) + \text{sg}_\xi(\mathcal{L}_G) \rightarrow \min_{\theta, \xi}, \quad (9)$$

where x denotes a sample from dataset, \hat{x} is the outputting prediction, and M denotes the circular masks to synthesized fisheye images. sg_k stops gradient w.r.t. k , and \mathcal{L}_{adv} is the joint adversarial loss which needs to be optimized.

Third, L_{rec} measures the L1 distance between the outpainted image generated by FishDreamer and the ground truth, which can be calculated via the following equation:

$$\mathcal{L}_{rec} = \|\hat{x} - x\|_1. \quad (10)$$

Then, L_{fm} [57] is leveraged to denote a discriminated-based perceptual loss, which stabilizes training and improves the performance.

Finally, for the semantic completion we employ further two additional loss functions, *i.e.*, a cross entropy loss \mathcal{L}_{ce} and a Lovász-softmax loss \mathcal{L}_{lovasz} [3].

The final FishDreamer loss \mathcal{L} becomes:

$$\mathcal{L} = \lambda_{adv} \cdot L_{adv} + \lambda_{hp} \cdot \mathcal{L}_{hp} + \lambda_{fm} \cdot \mathcal{L}_{fm} + \lambda_{rec} \cdot \mathcal{L}_{rec} + \lambda_{ce} \cdot \mathcal{L}_{ce} + \lambda_{lovasz} \cdot \mathcal{L}_{lovasz}, \quad (11)$$

which is the weighted sum of the above losses. We empirically set $\lambda_{adv}=20$, $\lambda_{hp}=60$, $\lambda_{fm}=200$, $\lambda_{rec}=20$, $\lambda_{ce}=30$, and $\lambda_{lovasz}=10$ in all the experiments.

Table 1. The training and validation sets splits of the derived Cityscapes-BF and KITTI360-BF.

Dataset	Train	Validation	Total
Cityscapes-BF	2,975	500	3,475
KITTI360-BF	9,800	2,455	12,255

4. Experiments

4.1. Datasets

To facilitate training and evaluation of fisheye semantic completion, we derive Cityscapes-BF and KITTI360-BF from the Cityscapes [7] and KITTI-360 [28] datasets. To achieve this, we conduct a radial distortion of perspective images via the following equation [26]:

$$\begin{aligned} x_d &= x_o(1 + k_1 r_f^2 + k_2 r_f^4 + k_3 r_f^6 + k_4 r_f^8 + \dots) \\ y_d &= y_o(1 + k_1 r_f^2 + k_2 r_f^4 + k_3 r_f^6 + k_4 r_f^8 + \dots), \end{aligned} \quad (12)$$

where $\mathbf{P}_o = (x_o, y_o)^T \in \mathbb{R}^{2 \times 1}$ is a pixel in the original pinhole image and $\mathbf{P}_d = (x_d, y_d)^T \in \mathbb{R}^{2 \times 1}$ is its corresponding pixel in distorted fisheye image. $[k_1, k_2, k_3, k_4, \dots]$ are the radial distortion parameters and r_f is the Euclidean distance between the distorted pixel and the distortion center $\mathbf{P}_c = (x_c, y_c)^T \in \mathbb{R}^{2 \times 1}$.

Following the previous work of FisheyeEX [25], we use the same fourth order polynomial model to apply a distortion on original images. Besides, we acquire a circular mask by applying a mask generator. It masks out a circular region of the ground-truth image, aiming to make our synthesized fisheye image like the natural fisheye images captured via a fisheye camera. Therefore, our datasets are composed of: complete-FoV fisheye images, complete-FoV fisheye semantic labels, and circular masks. The datasets distribution are shown in Fig. 5.

4.2. Implementation Details

FishDreamer was implemented in PyTorch and trained for fisheye semantic completion end-to-end on an NVIDIA RTX 3090 graphics card. Backbone weights are initialized from models pretrained on ImageNet [8]. We choose the

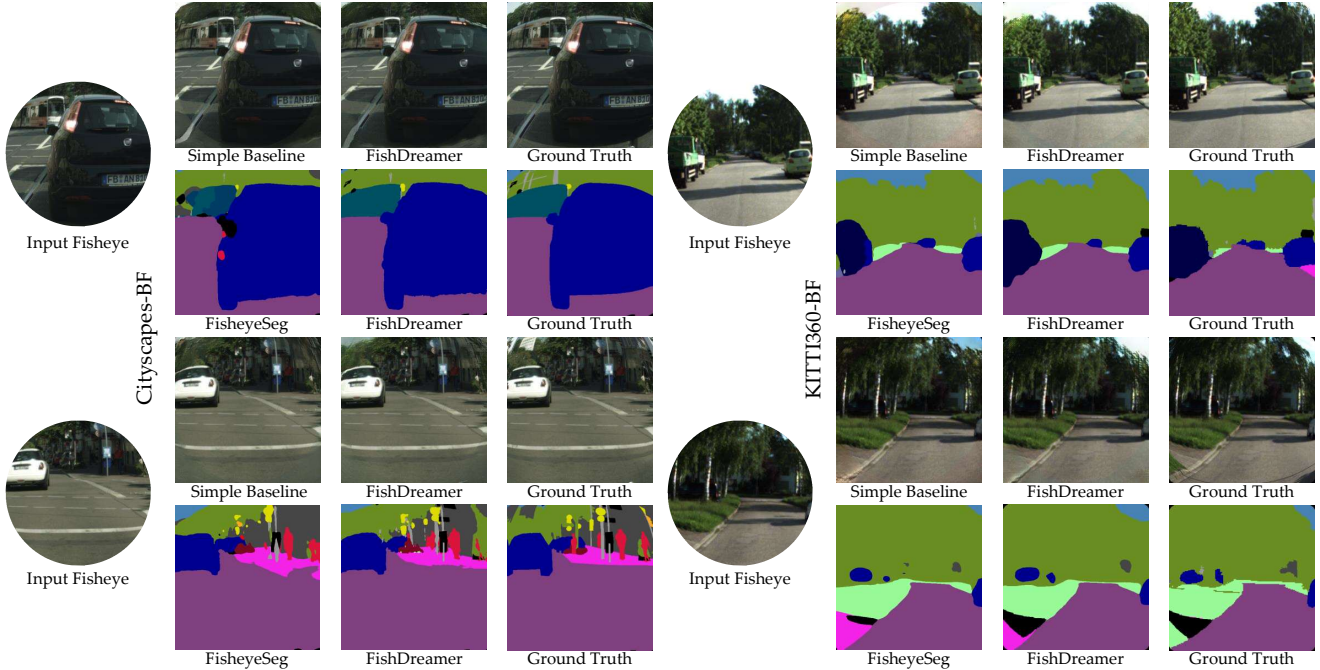


Figure 6. Qualitative comparisons with alternative methods on Cityscapes-BF and KITTI360-BF. We compare the proposed FishDreamer with our simple baseline since it already surpasses previous works [25] in terms of PSNR without the help of PCA. We also compare with the fisheye semantic segmentation method FisheyeSeg [70]. Best viewed in color.

Table 2. Quantitative comparison of fisheye semantic completion on Cityscapes-BF dataset. * indicates results from [25].

Dataset	Cityscapes-BF			
	Image Completion		Semantic Completion	
Method	PSNR \uparrow	SSIM \uparrow	FID \downarrow	mIoU \uparrow
SRN* [58]	17.71	0.81	169.42	<i>n.a.</i>
RK* [32]	21.79	0.87	136.66	<i>n.a.</i>
HiFill* [71]	22.27	0.89	109.89	<i>n.a.</i>
Boundless* [53]	23.54	0.90	64.26	<i>n.a.</i>
FisheyeEX* [25]	24.63	0.92	40.06	<i>n.a.</i>
FisheyeSeg [70]	<i>n.a.</i>	<i>n.a.</i>	<i>n.a.</i>	47.31
Swin-S + UPerNet	<i>n.a.</i>	<i>n.a.</i>	<i>n.a.</i>	53.65
SegFormer-B2 [60]	<i>n.a.</i>	<i>n.a.</i>	<i>n.a.</i>	53.80
Simple Baseline (Ours)	24.82	0.93	34.48	53.98
FishDreamer (Ours)	25.05	0.93	30.14	54.54

AdamW optimizer [35] with a learning rate of 2.5×10^{-4} , coefficients $\beta_1=0.9$, $\beta_2=0.999$, and weight decay $\eta=10^{-2}$. We use batch size of 8 and train our model for 50/70 epochs for the ablation experiments and the final model experiments respectively. We adopt Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Measure (SSIM), and Fréchet Inception Distance (FID) to evaluate the extrapolation performance, while mean Intersection over Union (mIoU) is used as the semantic completion metric.

4.3. Results

Cityscapes-BF results. In Table 2, we conduct comparison between the state-of-the-art methods for image completion, *e.g.*, FisheyeEX [25], those for semantic segmentation, *e.g.*, SegFormer [60], and the proposed FishDreamer approach.

Table 3. Quantitative comparison of fisheye semantic completion on KITTI360-BF dataset. * indicates results from [25].

Dataset	KITTI360-BF			
	Image Completion		Semantic Completion	
Method	PSNR \uparrow	SSIM \uparrow	FID \downarrow	mIoU \uparrow
SRN* [58]	18.25	0.79	143.90	<i>n.a.</i>
RK* [32]	20.13	0.82	102.77	<i>n.a.</i>
HiFill* [71]	20.10	0.83	82.61	<i>n.a.</i>
Boundless* [53]	21.52	0.87	53.17	<i>n.a.</i>
FisheyeEX* [25]	22.31	0.90	34.68	<i>n.a.</i>
FisheyeSeg [70]	<i>n.a.</i>	<i>n.a.</i>	<i>n.a.</i>	39.08
Swin-S + UPerNet	<i>n.a.</i>	<i>n.a.</i>	<i>n.a.</i>	40.00
SegFormer-B2 [60]	<i>n.a.</i>	<i>n.a.</i>	<i>n.a.</i>	41.19
Simple Baseline (Ours)	22.38	0.90	30.23	42.24
FishDreamer (Ours)	22.51	0.91	27.89	43.57

Compared to the previous FisheyeEX, FishDreamer obtains better results on the sub-task of image completion, which are respective 25.05, 0.93, and 30.14 in PSNR, SSIM, and FID. On the sub-task of semantic completion, FishDreamer obtains the best score of 54.54% in mIoU, yielding a large performance boost in +7.23% as compared to the previous fisheye image semantic segmentation model FisheyeSeg [70]. Besides, compared to the methods for general image semantic segmentation, such as Swin [33] and SegFormer [60], our FishDreamer model also yields considerable improvement. As the completion involves severe distortions and demands inferring semantics beyond the FoV, the segmentation transformers deliver clearly lower scores compared to their performances on standard segmentation benchmarks. Yet, the state-of-the-art performance achieved

Table 4. Ablation of different backbones.

Backbone	PSNR	SSIM	mIoU	#Params(M)
Conformer-T [40]	24.34	0.9227	47.36	<u>29.0</u>
Conformer-S [40]	24.57	0.9237	49.81	45.0
MiT-B0 [60]	23.28	0.9203	44.83	9.3
MiT-B2 [60]	24.12	0.9229	50.72	30.9
Swin-T [33]	24.21	0.9167	50.96	34.9
Swin-S [33]	<u>24.46</u>	0.9224	54.01	56.2

in both tasks of FishDreamer demonstrates that outpainting the content outside the fisheye camera’s FoV and completing the semantics of objects are conjugated. In other words, by using a single end-to-end model for both problems we can effectively leverage this complementary information, yielding clear benefits for both tasks.

KITTI360-BF results. As shown in Table 3, the results on the KITTI360-BF dataset are compared among fisheye image completion methods, semantic completion methods, and our two fisheye semantic completion approaches. Compared to image completion methods like Boundless [53] and FisheyeEX [25], our simple baseline achieves the second-best result (22.38) with significant improvement. Moreover, our FishDreamer model attains the best results in the image completion sub-task, with respective scores of 22.51, 0.91, and 27.89 in PSNR, SSIM, and FID. In addition to image completion, our FishDreamer achieves the best semantic completion performance with an mIoU of 43.57. These results and improvements further demonstrate the promising performance of our proposed method, which effectively couples both sub-tasks in an end-to-end manner.

4.4. Ablation Studies

Analysis of the backbones. To investigate the effect of model backbones, we perform ablation study of FishDreamer with three different methods, including Swin [33], MiT [60], and Conformer [40]. As shown in Table 4, each method has two model scales. The best and the second best results are marked with bold and underline, respectively. Our method based on Conformer [40] models have better performance on the sub-task of fisheye image completion, and the model based on Conformer-S achieves respective 24.57 and 0.9237 scores in PSNR and SSIM, but obtains sub-optimal performance in semantic completion, yielding only 49.81 in mIoU. Compared to the Conformer models, based on MiT-B0 and -B2 [60] backbones that are specific for semantic segmentation, our method achieves better results on the sub-task of semantic completion with 44.83 and 50.72 scores in mIoU. The MiT-based models have a smaller number of parameters (9.3M and 30.9M), however, the performance on the fisheye image completion is lower as compared to the ones using Conformer counterparts. To achieve a balance between the two sub-tasks, the Swin-based [33] backbone strikes a good balance between

Table 5. Analysis of Polar-aware Cross Attention.

Polar Mask	Direction	mIoU	PSNR	SSIM
<i>w/o</i>	S2P	53.73	<u>24.93</u>	0.9246
	P2S	<u>54.17</u>	24.56	0.9215
	Bi-direction	53.89	24.82	0.9240
2	Bi-direction	54.10	<u>24.93</u>	<u>0.9249</u>
4	Bi-direction	54.21	25.01	0.9257
8	Bi-direction	53.84	24.82	0.9242

the fisheye image completion performance and the semantic completion quality. Our method based on Swin-S backbone obtains the best semantic completion result with 54.01 in mIoU, while it provides the second best result on fisheye image completion with 24.46 in PSNR. This result aligns with our observation that a backbone with superior semantic completion capabilities can provide complementary advantages for the image completion sub-task.

Analysis of the Polar-aware Cross Attention (PCA). The PCA mechanism is vital for the fisheye semantic completion task. To examine the impact of mask selection and direction, we perform an ablation study of PCA in Table 5. Without using the polar mask, the three ways of semantic-to-outpainting (S2P), outpainting-to-semantic (P2S), and Bi-direction achieve respective 53.73, 54.17, and 53.89 in mIoU of semantic completion task, and 24.93, 24.56, and 24.82 in image completion PSNR. When using the polar mask and Bi-direction method, we further ablate the mask generation with different mask numbers in {2, 4, 8}. As shown in Table 5, our PCA module is robust to different mask numbers, since each of them obtains comparable performance. Nonetheless, we found that using 4 polar masks could provide better results on both sub tasks, yielding 54.21, 25.01, and 0.9257 in mIoU, PSNR, and SSIM, respectively. This analysis demonstrates that our proposed PCA module is effective in simultaneously addressing semantic understanding and image completion.

5. Conclusion

In this paper, we look into fisheye semantic completion, a novel task addressing the extended field of view perception in terms of simultaneous image sensing and semantic understanding. To tackle this challenge, we propose FishDreamer to intertwine image outpainting and segmentation via polar-aware cross attention, which guides outpainting with extended semantic contextual information in the annular dimension. We establish Cityscapes-BF and KITTI360-BF benchmarks to assess the effectiveness of driving scene understanding beyond fisheye vision. Extensive experiments demonstrate the effectiveness of the proposed polar-aware cross attention and FishDreamer over its counterparts.

In this future, we intend to introduce conditional diffusion models to enhance the completion and investigate the generalization of FishDreamer on real fisheye images.

References

- [1] Ola Ahmad and Freddy Lecue. FisheyeHDK: Hyperbolic deformable kernel learning for ultra-wide field-of-view image recognition. In *AAAI*, 2022. 3
- [2] Bruno Arsenali, Prashanth Viswanath, and Jelena Novosel. RotinvMTL: Rotation invariant MultiNet on fisheye images for autonomous driving applications. In *ICCVW*, 2019. 3
- [3] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, 2018. 6
- [4] Gregor Blott, Masato Takami, and Christian Heipke. Semantic segmentation of fisheye images. In *ECCVW*, 2018. 2
- [5] Vincent Cartillier, Zhile Ren, Neha Jain, Stefan Lee, Irfan Essa, and Dhruv Batra. Semantic MapNet: Building allocentric semantic maps and representations from egocentric views. In *AAAI*, 2021. 1
- [6] Ramchandra Cheke, Ganesh Sistu, Ciarán Eising, Pepijn van de Ven, Varun Ravi Kumar, and Senthil Yogamani. FisheyePixPro: Self-supervised pretraining using fisheye images for semantic segmentation. In *AVM*, 2022. 3
- [7] 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. 2, 6
- [8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009. 6
- [9] 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. *T-ITS*, 2020. 3
- [10] 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. 2
- [11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 5
- [12] Ciarán Eising, Jonathan Horgan, and Senthil Yogamani. Near-field perception for low-speed vehicle automation using surround-view fisheye cameras. *T-ITS*, 2022. 3
- [13] Shaohua Gao, Kailun Yang, Hao Shi, Kaiwei Wang, and Jian Bai. Review on panoramic imaging and its applications in scene understanding. *TIM*, 2022. 1
- [14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 2020. 2
- [15] Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, and Jose J. Guerrero. What’s in my room? object recognition on indoor panoramic images. In *ICRA*, 2020. 1
- [16] Xing Hu, Yi An, Cheng Shao, and Huosheng Hu. Distortion convolution module for semantic segmentation of panoramic images based on the image-forming principle. *TIM*, 2022. 3
- [17] Sujin Jang, Joochan Na, and Dokwan Oh. DaDA: Distortion-aware domain adaptation for unsupervised semantic segmentation. In *NeurIPS*, 2022. 3
- [18] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Towards complete surrounding understanding via unsupervised contrastive learning. In *IV*, 2021. 3
- [19] Alexander Jaus, Kailun Yang, and Rainer Stiefelhagen. Panoramic panoptic segmentation: Insights into surrounding parsing for mobile agents via unsupervised contrastive learning. *T-ITS*, 2023. 3
- [20] Deukhyeon Kim, Jinwook Choi, Hunjae Yoo, Ukil Yang, and Kwanghoon Sohn. Rear obstacle detection system with fish-eye stereo camera using HCT. *Expert Systems with Applications*, 2015. 1
- [21] Jihyun Kim, Somi Jeong, and Kwanghoon Sohn. PASTS: Toward effective distilling transformer for panoramic semantic segmentation. In *ICIP*, 2022. 3
- [22] Varun Ravi Kumar, Ciarán Eising, Christian Witt, and Senthil Yogamani. Surround-view fisheye camera perception for automated driving: Overview, survey & challenges. *T-ITS*, 2023. 1
- [23] Varun Ravi Kumar, Marvin Klingner, Senthil Yogamani, Stefan Milz, Tim Fingscheidt, and Patrick Mader. SynDistNet: Self-supervised monocular fisheye camera distance estimation synergized with semantic segmentation for autonomous driving. In *WACV*, 2021. 3
- [24] Varun Ravi Kumar, Senthil Yogamani, Hazem Rashed, Ganesh Sitsu, Christian Witt, Isabelle Leang, Stefan Milz, and Patrick Mäder. OmniDet: Surround view cameras based multi-task visual perception network for autonomous driving. *RA-L*, 2021. 3
- [25] Kang Liao, Chunyu Lin, Yunchao Wei, and Yao Zhao. FisheyeEX: Polar outpainting for extending the FoV of fish-eye lens. *arXiv preprint arXiv:2206.05844*, 2022. 2, 6, 7, 8
- [26] Kang Liao, Chunyu Lin, Yao Zhao, and Mai Xu. Model-free distortion rectification framework bridged by distortion distribution map. *TIP*, 2020. 6
- [27] Kang Liao, Lang Nie, Chunyu Lin, Zishuo Zheng, and Yao Zhao. RecRecNet: Rectangling rectified wide-angle images by thin-plate spline model and DoF-based curriculum learning. *arXiv preprint arXiv:2301.01661*, 2023. 2
- [28] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2D and 3D. *TPAMI*, 2022. 2, 6
- [29] Ailiang Lin, Bingzhi Chen, Jiayu Xu, Zheng Zhang, Guangming Lu, and David Zhang. DS-TransUNet: Dual swin transformer u-net for medical image segmentation. *TIM*, 2022. 3
- [30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

- [31] Yih-Shiuan Lin, Chien-Chung Chen, and Mark W. Greenlee. Neural correlates of lateral modulation and perceptual filling-in in center-surround radial sinusoidal gratings: an fMRI study. *Scientific Reports*, 2022. 1
- [32] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *ECCV*, 2020. 7
- [33] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 7, 8
- [34] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, 2022. 3
- [35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 7
- [36] Chaoxiang Ma, Jiaming Zhang, Kailun Yang, Alina Roitberg, and Rainer Stiefelhagen. DensePASS: Dense panoramic semantic segmentation via unsupervised domain adaptation with attention-augmented context exchange. In *ITSC*, 2021. 3
- [37] Jieru Mei, Alex Zihao Zhu, Xinchun Yan, Hang Yan, Siyuan Qiao, Liang-Chieh Chen, and Henrik Kretzschmar. Waymo open dataset: Panoramic video panoptic segmentation. In *ECCV*, 2022. 3
- [38] Semih Orhan and Yalin Bastanlar. Semantic segmentation of outdoor panoramic images. *SIVP*, 2022. 3
- [39] Kunyu Peng, Alina Roitberg, Kailun Yang, Jiaming Zhang, and Rainer Stiefelhagen. TransDARC: Transformer-based driver activity recognition with latent space feature calibration. In *IROS*, 2022. 3
- [40] Zhiliang Peng, Wei Huang, Shanzhi Gu, Lingxi Xie, Yaowei Wang, Jianbin Jiao, and Qixiang Ye. Conformer: Local features coupling global representations for visual recognition. In *ICCV*, 2021. 8
- [41] Luiz Pessoa and Peter De Weerd. *Filling-in: From perceptual completion to cortical reorganization*. Oxford University Press, 2003. 1
- [42] Clément Ployout, Ola Ahmad, Freddy Lecue, and Farida Cheriet. Adaptable deformable convolutions for semantic segmentation of fisheye images in autonomous driving systems. *arXiv preprint arXiv:2102.10191*, 2021. 3
- [43] Ye Qiang Qian, Ming Yang, and John M. Dolan. Survey on fish-eye cameras and their applications in intelligent vehicles. *T-ITS*, 2022. 1
- [44] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 4
- [45] Mark Sabini and Gili Rusak. Painting outside the box: Image outpainting with GANs. *arXiv preprint arXiv:1808.08483*, 2018. 2
- [46] Alvaro Sáez, Luis M. Bergasa, Eduardo Romeral, Elena López, Rafael Barea, and Rafael Sanz. CNN-based fisheye image real-time semantic segmentation. In *IV*, 2018. 2
- [47] Ahmed Rida Sekkat, Yohan Dupuis, Paul Honeine, and Pascal Vasseur. A comparative study of semantic segmentation of omnidirectional images from a motorcycle perspective. *Scientific Reports*, 2022. 2
- [48] Ahmed Rida Sekkat, Yohan Dupuis, Pascal Vasseur, and Paul Honeine. The OmniScape dataset. In *ICRA*, 2020. 1
- [49] Qi Shan, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M. Seitz. Photo uncrop. In *ECCV*, 2014. 2
- [50] Hao Shi, Qi Jiang, Kailun Yang, Xiaoting Yin, and Kaiwei Wang. FlowLens: Seeing beyond the FoV via flow-guided clip-recurrent transformer. *arXiv preprint arXiv:2211.11293*, 2022. 2
- [51] Yongjie Shi, Xianghua Ying, and Hongbin Zha. Unsupervised domain adaptation for semantic segmentation of urban street scenes reflected by convex mirrors. *T-ITS*, 2022. 3
- [52] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, 2022. 5
- [53] Piotr Teterwak, Aaron Sarna, Dilip Krishnan, Aaron Maschinot, David Belanger, Ce Liu, and William T. Freeman. Boundless: Generative adversarial networks for image extension. In *ICCV*, 2019. 2, 7, 8
- [54] Mohamed Thioune, Sanaa Chafik, Ankur Mahtani, Olivier Laurendin, and Safia Boudra. FPDM: Fisheye panoptic segmentation dataset for door monitoring. In *AVSS*, 2022. 3
- [55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3
- [56] Miao Wang, Yu-Kun Lai, Yuan Liang, Ralph R Martin, and Shi-Min Hu. BiggerPicture: data-driven image extrapolation using graph matching. *ACM Transactions on Graphics*, 2014. 2
- [57] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional GANs. In *CVPR*, 2018. 6
- [58] Yi Wang, Xin Tao, Xiaoyong Shen, and Jiaya Jia. Wide-context semantic image extrapolation. In *CVPR*, 2019. 2, 7
- [59] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *ECCV*, 2018. 3, 4
- [60] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 7, 8
- [61] Yuanyou Xu, Kaiwei Wang, Kailun Yang, Dongming Sun, and Jia Fu. Semantic segmentation of panoramic images using a synthetic dataset. In *SPIE*, 2019. 3
- [62] Kailun Yang, Xinxin Hu, Luis M. Bergasa, Eduardo Romera, Xiao Huang, Dongming Sun, and Kaiwei Wang. Can we PASS beyond the field of view? Panoramic annular semantic segmentation for real-world surrounding perception. In *IV*, 2019. 3
- [63] Kailun Yang, Xinxin Hu, Luis M. Bergasa, Eduardo Romera, and Kaiwei Wang. PASS: Panoramic annular semantic segmentation. *T-ITS*, 2020. 3

- [64] Kailun Yang, Xinxin Hu, Hao Chen, Kaite Xiang, Kaiwei Wang, and Rainer Stiefelwagen. DS-PASS: Detail-sensitive panoramic annular semantic segmentation through swaftnet for surrounding sensing. In *IV*, 2020. 3
- [65] Kailun Yang, Xinxin Hu, Yicheng Fang, Kaiwei Wang, and Rainer Stiefelwagen. Omnisupervised omnidirectional semantic segmentation. *T-ITS*, 2022. 3
- [66] Kailun Yang, Xinxin Hu, and Rainer Stiefelwagen. Is context-aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild. *TIP*, 2021. 3
- [67] Kailun Yang, Jiaming Zhang, Simon Reiß, Xinxin Hu, and Rainer Stiefelwagen. Capturing omni-range context for omnidirectional segmentation. In *CVPR*, 2021. 3
- [68] Shangrong Yang, Chunyu Lin, Kang Liao, and Yao Zhao. FishFormer: Annulus slicing-based transformer for fisheye rectification with efficacy domain exploration. *arXiv preprint arXiv:2207.01925*, 2022. 5
- [69] Kai Yao, Penglei Gao, Xi Yang, Jie Sun, Rui Zhang, and Kaizhu Huang. Outpainting by queries. In *ECCV*, 2022. 2
- [70] Yaozu Ye, Kailun Yang, Kaite Xiang, Juan Wang, and Kaiwei Wang. Universal semantic segmentation for fisheye urban driving images. In *SMC*, 2020. 2, 7
- [71] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *CVPR*, 2020. 7
- [72] 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. 1
- [73] Jiaming Zhang, Chaoxiang Ma, Kailun Yang, Alina Roitberg, Kunyu Peng, and Rainer Stiefelwagen. Transfer beyond the field of view: Dense panoramic semantic segmentation via unsupervised domain adaptation. *T-ITS*, 2022. 3
- [74] Jiaming Zhang, Kailun Yang, Chaoxiang Ma, Simon Reiß, Kunyu Peng, and Rainer Stiefelwagen. Bending reality: Distortion-aware transformers for adapting to panoramic semantic segmentation. In *CVPR*, 2022. 3
- [75] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Kaiwei Wang, and Rainer Stiefelwagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *arXiv preprint arXiv:2207.11860*, 2022. 3
- [76] Yinda Zhang, Jianxiong Xiao, James Hays, and Ping Tan. FrameBreak: Dramatic image extrapolation by guided shift-maps. In *CVPR*, 2013. 2
- [77] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 4
- [78] Zishuo Zheng, Chunyu Lin, Lang Nie, Kang Liao, Zhijie Shen, and Yao Zhao. Complementary bi-directional feature compression for indoor 360° semantic segmentation with self-distillation. In *WACV*, 2023. 3