DarSwin-Unet: Distortion Aware Architecture

Akshaya Athwale*[1], Ichrak Shili*[1], Émile Bergeron[1], Ola Ahmad[2], Jean-François Lalonde[1]
[1]Université Laval [2]Thales CortAIx Labs Canada

Abstract

Wide angle fisheye images are becoming increasingly common for perception tasks in applications such as robotics, security, and mobility (e.g. drones, avionics). However, current models often either ignore the distortions in wide angle images or are not suitable to perform pixel-level tasks. In this paper, we present an encoder-decoder model based on a radial transformer architecture that adapts to distortions in wide angle lenses by leveraging the physical characteristics defined by the radial distortion profile. In contrast to the original model, which only performs classification tasks, we introduce a U-Net architecture, DarSwin-Unet, designed for pixel level tasks. Furthermore, we propose a novel strategy that minimizes sparsity when sampling the image for creating its input tokens. Our approach enhances the model capability to handle pixel-level tasks in wide angle fisheye images, making it more effective for real-world applications. Compared to other baselines, DarSwin-Unet achieves the best results across different datasets, with significant gains when trained on bounded levels of distortions (very low, low, medium, and high) and tested on all, including out-of-distribution distortions. We demonstrate its performance on depth estimation and show through extensive experiments that DarSwin-Unet can perform zero-shot adaptation to unseen distortions of different wide angle lenses. The code and models are publicly available at https://lvsn.github.io/darswin-unet/.

1. Introduction

[1] Many areas in computer vision, such as security [17], augmented reality (AR) [40], healthcare, and particularly autonomous vehicles [10, 58], utilize wide angle lenses because they minimize costs by requiring fewer cameras to capture a 360° scene due to their increased field of view.

However, this cost benefit comes with a drawback: images captured by wide angle lenses exhibit significant distortion because the projection model is no longer perspective. Straight lines in the real world appear curved in the image, and the geometry of objects changes as a function of their location in the image. The majority of CNN-based models have an implicit bias towards perspective images—indeed, the distortions in wide angle lenses break the translational equivariance of CNNs, limiting their applicability. The diversity in wide angle lens distortions further exacerbates this problem: a method trained on a specific lens distortion does not generalize well when evaluated on another lens with different distortion—one must therefore repeat the entire data collection, training procedure, etc. on such a new lens.

One popular strategy to improve generalization when tested on another lens is canceling the distortion effect by warping the input image back to the perspective projection model. A wide array of such methods, ranging from classical [5, 15, 34, 37, 60] to deep learning [50, 56], have been proposed to train and test on the undistorted image. Unfortunately, canceling the effect of the distortion of wide angle images tends to create severely stretched images. It can also restrict the maximum field of view since, in the limit, a point at 90° azimuth projects at infinity, but reducing the maximum field of view defeats the purpose of using a wide angle lens in the first place. Other projections are also possible (e.g., cylindrical [35], or piecewise linear [58]), but these also tend to create unwanted distortions or suffer from resolution loss. Some methods like [1, 36] use deformable convolutions [9, 61] to reason on wide angle image without undistorting them. Here, convolution kernels adapt to the lens distortion of the given image during training. However, these methods tend to overfit to the wide angle lens distortion present at training; hence, they cannot generalize over unseen lens distortion at test time. Transformer-based architectures [2, 7] are also used to reason directly on the wide angle image, even evaluating the generalization performance to other distortions in DarSwin [2]. However, DarSwin was only demonstrated for classification, and HealSwin [7] could not adapt to other lenses at test time. To this date, it is not clear whether distortion-aware architectures can be trained for pixel-level tasks with zero-shot generalization to other distortion lenses at test time without fine-tuning.

In this work, we present a robust solution to that very problem. In particular we present an encoder-decoder architecture named DarSwin-Unet, which leverages DarSwin [2] as the encoder. While this strategy is effective, we observe that

[1]*Authors contributed equally

the image sampling pattern proposed in [2] creates sparsity issues which negatively affect the performance of pixel-level tasks such as depth estimation. To address this issue, we propose a novel pixel sampling method that mitigates the aforementioned sparsity problem, thereby significantly improving depth estimation performance. We present experiments on the depth estimation task, which show that DarSwin-Unet is much more robust to changes in lens distortions at test time than all of the compared baselines, including Swin-Unet [6] and Swin-UPerNet [32, 49] trained on both distorted and undistorted images, and DAT-UPerNet [47, 49] trained on distorted images.

In short, we make the following key contributions:

- a novel encoder-decoder distortion-aware architecture, named DarSwin-Unet, suitable for pixel-level tasks, which adapts to the distortion in wide angle images in a zero-shot manner at test time, without fine-tuning;
- a new pixel sampling scheme which limits the sample sparsity problem when dealing with images of drastically different distortions;
- extensive experiments on depth estimation showing the superiority of DarSwin-Unet at adapting to novel distortion profiles at test time.

2. Related work

Distortion correction Wide angle cameras are increasingly used in various computer vision applications, including visual perception [24] and autonomous vehicle cameras [19, 29, 58]. However, their adoption has been relatively recent [35, 38, 39, 55, 58] due to the distortion present in their images. Earlier methods primarily focused on correcting this distortion [11, 14, 16, 27, 28, 50, 52, 54, 56, 60]. Some approaches [27, 28, 53] use distortion parameters to assess distortion density per pixel and subsequently correct it. However, such correction processes can introduce artifacts like stretching [58], leading to performance degradation.

In contrast, our approach builds upon DarSwin [2], which directly utilizes distortion parameters to reason about wide angle images, avoiding the pitfalls associated with traditional distortion correction methods. This approach is crucial as it addresses distortions inherent in wide angle lenses and aligns with the needs of modern computer vision tasks.

Convolution-based approaches. CNNs [18, 30, 43] are highly effective for processing images with no distortion due to their inherent bias towards natural image characteristics, such as translational equivariance [4]. Methods like [38, 39, 44] try to adapt CNNs on fisheye images for tasks such as object detection. However, the distortion caused by wide angle images breaks this symmetry, which reduces the performance of CNNs. Methods like [21, 22, 25, 51] use self-supervised learning combined with techniques like distillation or multi-task learning to have a better understanding of distortion. Deformable convolutions [9, 61] offer flexibility by learning kernel deformations, though at a higher computational cost. Recent studies [1, 10, 36, 46] use deformable CNNs to handle fisheye distortion. In contrast, Our network builds on DarSwin [2], which leverages attention using a lens distortion profile instead of convolutions. However, unlike DarSwin's encoder-only design, our network, DarSwin-Unet, introduces an encoder-decoder architecture.

Hybrid-network based approaches. Some methods leverage properties from both self-attention and convolutions and build hybrid networks. Methods like [26] use hybrid networks and try to leverage the geometric property of fisheye images (i.e., the orthogonal placement of objects) and propose a new representation of fisheye road scenes, invariant to the camera viewing direction. Shi et al. [41] leverages the radial nature of distortion by including polar cross attention for inpainting, but unlike DarSwin-Unet, they do not use the lens information in their network. Similar to our method, [20, 21, 23, 25] propose a camera-aware depth estimation network to handle the severe distortion of fisheye cameras: [25] encode the camera intrinsic parameters as a tensor; and [23] propose a self-supervised depth estimation method which relies on the lens distortion parameter for forward and back-projection functions. Both these methods use distortion parameters as a part of the input or training process, but they rely on convolutions whose generalization capabilities are limited due to the translational invariance assumption being broken in wide angle images. Indeed, the network weights each pixel and its corresponding lens distortion prior equally regardless of the severity of distortion of the wide angle image, which affects the ability to generalize to a variety of lens distortion. Hence, [25] shows generalization to lens distortion closer to training distortion, and [23] does not show any generalization results.

Vision transformer based approaches. Vision transformers (ViT) [12] use self-attention mechanisms [45] computed on image patches rather than performing convolutions. Unlike CNNs, a ViT does not have a fixed geometric structure in its architecture: any extra structure is given via positional encoding. More recently, the Swin transformer architecture [32] incorporates a multi-scale strategy with window-based attention. Later, the Deformable Attention Transformer (DAT) [47] adopts the concept of deformable CNNs [9, 61] to enhance transformer adaptability. [59] proposes a distortion-aware architecture using a transformer network, but the network is limited to a fixed equirectangular distortion. Recently, methods very similar to our work, such as [2, 7] use the Swin transformer [32] as their base network. On one hand, [7] reasons on wide angle images by assuming a spherical projection model (sec. 3) and using a Healpix grid on sphere, unlike the Cartesian grid in

the original Swin transformer. However, it does not offer generalization capabilities to unseen lenses at test time. DarSwin [2], on the other hand, uses radial patches instead of the Cartesian grid in the Swin transformer network. It embeds the lens distortion parameter into the network (see sec. 3 for more details), to generalize the model's performance on unseen lens distortion at test time. However, DarSwin conducts its experiments only on the image classification task. In contrast, our proposed DarSwin-Unet, extends [2] to an encoder-decoder architecture to perform pixel-level tasks, making it more effective for real-world applications.

3. Background: Distortion-aware Radial Swin transformer (DarSwin)

This section briefly summarizes DarSwin [2], a distortion-aware radial patch-based encoder built on the Swin Transformer [32]. DarSwin adapts to wide angle distortions by dividing images into radial patches based on the lens distortion profile, as shown in fig. 1.

Architecture overview. The first layer of DarSwin divides the image into radial patches by defining the number of samples along radius N_r and azimuth N_φ. Samples along azimuth are obtained directly by dividing the angular dimension of the polar representation of the image into N_φ equal partitions as shown in fig. 2. Samples along the radius are obtained according to the lens curve ($r_d = \mathcal{P}(\theta)$) after dividing the axis along the incident angle θ into N_r equal partitions and sampling the radial value from the curve, as shown in fig. 1. Moreover, the examples in fig. 1 show that this partitioning strategy allows DarSwin to adapt to any lens, knowing its distortion curve, by changing the patch size. A CNN is then used to linearly embed these patches. However, since the patch sizes are different and the input to the CNN must have the same dimension, a fixed set of points are sampled for each patch as shown in fig. 2. After linear embedding, tokens are arranged in the polar format $N_r \times N_\varphi$, and are fed into the DarSwin self-attention blocks, which perform window-based self-attention. A set of non-overlapping shifted windows is defined using the patches along the azimuth dimension (N_φ), while shifts are obtained by displacing the windows along the azimuth. Finally, a downsampling step is used to reduce the spatial resolution by merging four angular patches along the azimuth prior to the next self-attention block. It is worth noting that DarSwin uses an angular relative positional encoding technique to capture the relative information between the produced radial tokens in its attention layers.

The original paper [2] evaluated DarSwin on synthetically distorted ImageNet (using the Unified camera model) and demostrated its ability to adapt to new distortion curves in a zero-shot test setting.

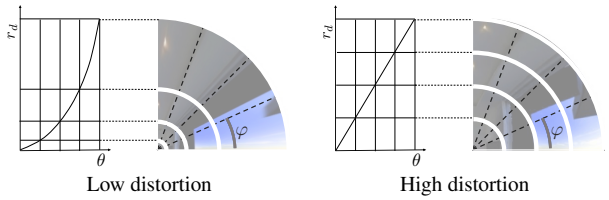


Figure 1. Radial divisions adapt to the lens distortion; here, we show low (left) and high (right) distortion for illustration purposes. DarSwin [2] separates radial patches equally along θ and determines the corresponding radius on the image plane according to the (known) lens distortion curve $r_d = \mathcal{P}(\theta)$.

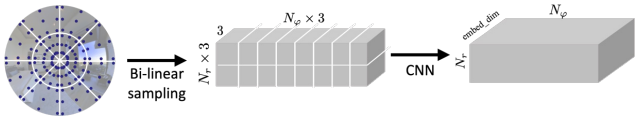


Figure 2. For illustration, the wide angle image is divided into 16 patches ($N_r = 2$ and $N_\varphi = 8$) along radius and azimuth. Nine samples are defined per patch: 3 samples along the radius and 3 samples along the azimuth. The image is bilinearly sampled and arranged in radial-azimuth format. This feature map is passed through CNN to embed each patch to get a feature map of dimension $N_r \times N_\varphi \times$ embed-dim.

Unified camera model. The Unified camera model [3, 33] describes the radial distortion by a *single*, *bounded* parameter $\xi \in [0, 1]$[2]. It projects the world point to the image as follows

$$r_d = \mathcal{P}(\theta) = \frac{f \cos\theta}{\xi + \sin\theta}, \tag{1}$$

focal length, and ξ the distortion parameter. We use this model for its flexibility in generating diverse distortion profiles with a single parameter ξ (fig. 7) and its analytically invertible mapping, though our approach is not restricted to this projection model.

4. Methodology

Inspired by the original DarSwin work [2], we propose two main contributions. First, we extend the DarSwin encoder-only architecture to a full encoder-decoder architecture (sec. 4.1). Second, we observe that sampling along the incident angle θ yields suboptimal coverage, leading to sparsity issues that affect performance. Therefore, we design a novel sampling technique (sec. 4.2) aimed at reducing sparsity by minimizing the distance between samples, thus improving performance.

4.1. DarSwin-Unet architecture

Fig. 3 shows an overview of the architecture, which takes in a wide angle image and its distortion curve $\mathcal{P}(\theta)$ as input.

[2] ξ can be slightly greater than 1 for certain types of catadioptric cameras [57] but this is ignored here.

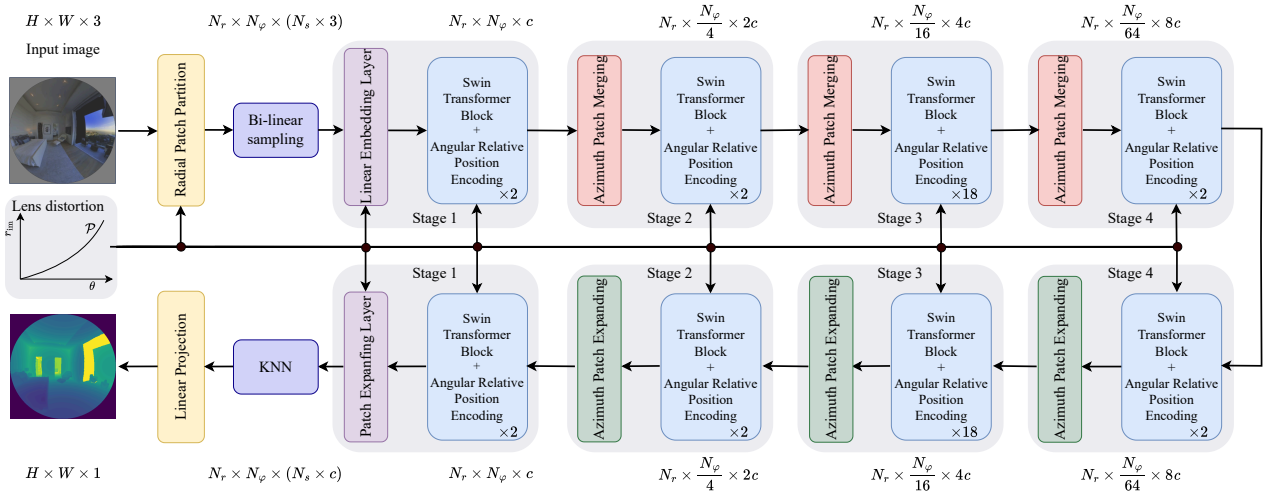


Figure 3. Overview of our distortion-aware transformer encoder-decoder architecture, DarSwin-Unet. It employs hierarchical layers of DarSwin transformer blocks [2] (top row) and replicates the structure in the decoder (similar to Swin-Unet [6]). To make the architecture adapt to lens distortion, the patch partition, linear embedding, patch merging, and patch expanding layers, all take the lens projection curve $\mathcal{P}(\theta)$ (c.f. sec. 3) as input. The k-NN layer is used to project the feature map from polar $(N_r \times N_\varphi)$ to cartesian space $H \times W$.

We propose a UNet architecture for our pixel-level model, where the encoder part is a DarSwin architecture, and a decoder part incorporating two novel components: an azimuth patch expanding layer for upsampling, and a k-NN layer to project the outputs into the Cartesian coordinates, providing pixel-level values as explained below.

Azimuth patch expanding layer. As explained in sec. 3, DarSwin uses an azimuth patch merging layer to downsample the feature map. The radial nature of DarSwin enables various possibilities when merging patches: merging along the radius, along the azimuth, or both. Here, the encoder employs azimuth patch merging, as it is found to perform best according to [2]. Consequently, we propose an azimuth patch expanding strategy.

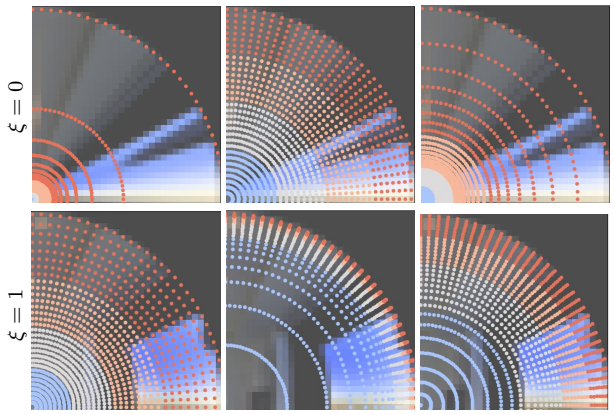
As in [6], we use an MLP for the expanding layer. We use this layer along the azimuth dimension to upsample by a $4\times$ factor. For example, consider the first (rightmost) patch expanding layer in fig. 3. The input feature map $(N_r \times \frac{N_\varphi}{64} \times 8c)$ is first given to an MLP layer to expand the feature dimension by $4\times$ to get $(N_r \times \frac{N_\varphi}{64} \times 32c)$ where N_r and N_φ are number of divisions along radius and azimuth respectively (c.f. sec. 3). The feature map is then rearranged to reduce the feature dimension and increase the resolution of the feature map along the azimuth dimension to obtain $(N_r \times \frac{N_\varphi}{16} \times 4c)$.

k-NN layer. Lastly, we employ a k-NN layer to map the polar feature map back to Cartesian coordinates. Each pixel coordinate in the image is associated with its k closest samples (we use $k = 4$), and their respective feature vectors are averaged. Since sample point locations are known, the k-NN layer is fixed and not trainable. The k-NN output (of dimensions $H \times W \times c$) is fed into the last linear projection layer to get the desired output for the required task.

4.2. Proposed sampling method

In the original DarSwin architecture, the input to the linear embedding layer (fig. 3) must have the same dimension. Therefore, a fixed set of points are sampled from the image for each patch. However, because the patch dimensions change according to the distortion, this can create sparse samples along the radius. Fig. 4 shows the samples obtained with two extreme distortions using the unified camera model with $\xi = 0$ and $\xi = 1$, respectively. Fig. 4-(a) shows how sampling according to the lens curve, $\mathcal{P}(\theta)$, used in [2], results in significant sparsity in the image, where many samples are missing across the radius, under the perspective projection ($\xi = 0$). Sampling according to another function of θ, for example $\mathcal{P}(\tan\theta)$ in fig. 4-(b), enhances the results but creates sparse samples when the distortion is very high (e.g., $\xi = 1$). Here, we are looking for another function of θ, named $g(\theta)$ which spreads out samples as uniformly as possible across a wide range of distortions (fig. 4-(c)). The lens functions are plotted in fig. 5. We observe that when the derivative (slope) of the lens function is high, the samples are spread far apart in the image.

More formally, we are looking for a strictly monotonic function g that has minimal derivative over its entire range when applied under both low and high distortion. As a proxy



(a) $r_d = \mathcal{P}(\theta)$ (b) $r_d = \mathcal{P}(\tan\theta)$ (c) $r_d = \mathcal{P}(g(\theta))$

Figure 4. Illustration of sampling (represented by colored dots) on a quadrant of an image taken from two different lenses ($\xi = 0$ (top row) and $\xi = 1$ (bottom row). The images is sampled according to the lens distortion curve $\mathcal{P}$ applied on different functions of θ: (a) θ, (b) $\tan\theta$, and (c) our novel $g(\theta)$. Observe how the first two options create large holes at either extreme values of ξ. In contrast, our proposed function offers a good compromise across a wide range of distortions.

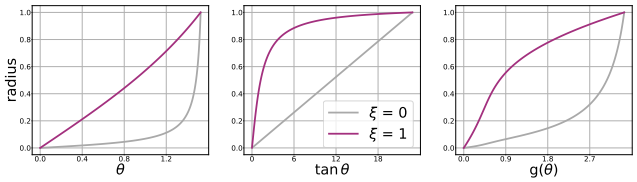


Figure 5. Lens distortion curves for least ($\xi = 0$) to most ($\xi = 1$) distorted using the unified camera model for illustration. We represent the same curves according to, from left to right, $\tan\theta$, our new $g(\theta)$, and θ. The high slopes present in both $\tan\theta$ and θ curves mean that samples will be spread far apart on the image plane. In contrast, our $g(\theta)$ offers a good compromise across the range of distortions.

for representing distortion, we again employ the unified camera model (c.f. sec. 3) and are looking for a function g that minimizes

$$\max_{\theta}\left(\frac{d\mathcal{P}(g(\theta))|_{\xi=0}}{d(g(\theta))}\right) + \max_{\theta}\left(\frac{d\mathcal{P}(g(\theta))|_{\xi=1}}{d(g(\theta))}\right). \quad (2)$$

We minimize the derivative only for low and high distortions, as the derivatives for all intermediate distortions lie between these two extremes (details in the supplementary). The function g is parameterized as a convex combination of two monotonic functions p_n and q_m.

$$g(\theta) = \lambda p_n(\theta) + (1-\lambda)q_m(\theta), \quad (3)$$
$$\text{with } p_n(\theta) = b\left(\frac{\theta}{a}\right)^n \text{ and } q_m(\theta) = 1 - \left(1 - \frac{\theta}{a}\right)^m.$$

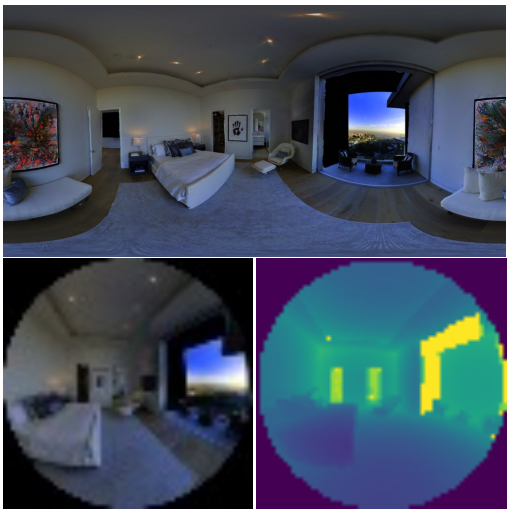


Figure 6. From a 360° panorama (top) from the Matterport3D dataset [8], we generate a wide angle image and its depth map (bottom) with a field of view 175°. Lens distortion is simulated with the uniform camera model (here, $\xi = 0.95$).

This ensures that the resulting curve is monotonic (see the supplementary for more details). We search the optimal parameters λ, m, n and b, that minimize the objective in eq. (2). For optimization, we perform an exhaustive search for $\lambda \in [0, 1]$ with 10 steps, $m \in [1, 20]$ with 60 steps, $n \in [0.5, 5]$ with 20 steps, and $b \in [2, 10]$ with 40 steps. We find that the values $\lambda = 0.777, m = 5.5084, n = 5.0, a = \frac{\text{FOV}}{2}, b = 4.1052$ gives us an optimal curves for both $\xi = 0$ and $\xi = 1$ as shown in fig. 5 (right). Using this optimal curve $g(\theta)$ for sampling reduces sparsity at both extreme cases (i.e., zero and maximal distortion levels) compared to the previous methods (fig. 4-(c)). In our experiments we sample 25 points along azimuth and 4 points along radius, in total 100 sample points per patch. We also analyze the performance on the depth estimation task using these three functions in supplementary material.

5. Depth estimation experiments

To evaluate the efficacy of DarSwin-Unet's generalization and robustness on unseen distortion profiles, we perform monocular depth estimation experiments using synthetically generated wide angle images using a panoramic dataset [8].

5.1. Datasets

Existing wide-angle depth estimation datasets, such as Woodscapes [58], lack the diverse distortion profiles required to evaluate our network's generalization. To address this, we generate synthetic wide-angle images by cropping 175° field-of-view images from Matterport3D panoramas [8] and simulating lens distortion using the unified camera model [3, 33] (see sec. 3), as illustrated in fig. 6.

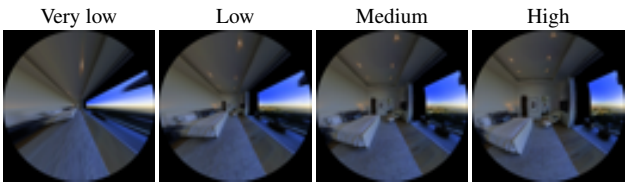
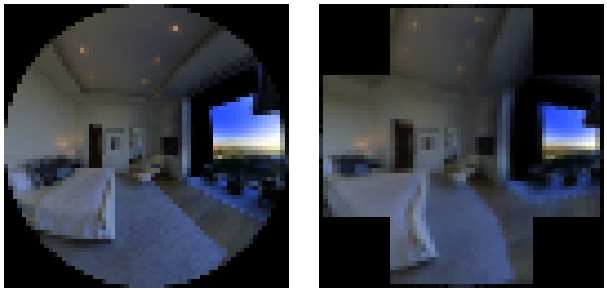


Figure 7. Visualization of a wide angle crop from a panorama with different distortions representing 4 different distortion levels From left to right: very low, low, medium, and high, used in four different training sets as explained below. The image is cropped from panorama with an original resolution of 512×1024, where the generated wide angle image is subsequently down-sampled to 64×64 after warping.



(a) Wide angle image (b) Piecewise perspective

Figure 8. (a) 175° field of view wide angle image and (b) corresponding undistorted image using piecewise linear correction (cubemap representation).

Training set. Similar to DarSwin [2], we generate four different training sets with different levels of distortion, defined by the distortion parameter ξ: "very low" ($\xi \in [0.0, 0.05]$), "low" ($\xi \in [0.2, 0.35]$), "medium" ($\xi \in [0.5, 0.7]$), and "high" ($\xi \in [0.85, 1.0]$) as illustrated in fig. 7.

Training images are synthetically generated from panoramas on the fly during training with distortion ξ sampled from their respective intervals, and the yaw viewing angle in the panorama is uniformly sampled in the $[0, 360^\circ]$ interval. Each of the four training sets (one for each distortion group) contains $9,180$ panoramas of original resolution of 512×1024, where the generated wide angle image is subsequently downsampled to 64×64 after warping.

Test set. To evaluate performance and zero-shot generalization to seen and unseen distortion profiles, we generate 20 test sets, each with a fixed distortion value uniformly sampled from $\xi \in [0, 1]$ in steps of 0.05. All test sets are created from the same $1,620$ test panoramas.

5.2. Baselines

We compare to the following baselines. First, we use Swin-Unet [6] and Swin-UPerNet (a Swin [32] encoder with UPerNet [48] decoder). We also compare with DAT [47], which leverages deformable attention in order to understand lens distortion for better robustness. Hence, we compare with DAT-UPerNet (a DAT [47] encoder with UPerNet decoder). Since these baselines do not have access to the lens distortion, as opposed to our proposed DarSwin-Unet, we also correct the distortion in the image and train the Swin baselines on this input. We dub these alternatives as Swin-Unet(undis) and Swin-UPerNet(undis).

Undistorting with piecewise perspective projection. Undistorting a 175° field-of-view wide angle image to a single perspective image will result in extremely severe stretching. Instead, we follow the piecewise perspective correction strategy in [58] and undistort the image to a partial cubemap, which is composed of 6 perspective faces of 90° field of view each, unrolled into an image and cropped to keep only the valid pixels. As shown in fig. 8, this preserves the entire field of view while minimizing stretching. As mentioned above, these images are used to train the Swin-Unet(undis) and Swin-UPerNet(undis) baselines.

5.3. Training details and evaluation metrics

All baselines have 1024 patch divisions on image size 64×64 with patch size 2×2 and window size 4×4 along the height and width. For DarSwin-Unet, we employ 16 divisions along the radius and 64 on the azimuth on an image with a total of 1024 divisions (to have the same number of patches as baselines). All encoders (Swin and DAT) are first pre-trained for classification on the distorted tiny-ImageNet dataset from [2]. Pre-trained encoders, along with their respective decoders, are fine-tuned on the depth estimation task. All methods are trained with the SGD optimizer with momentum 0.9 and weight decay 10^{-4} with a batch size of 8. We employ a polynomial learning rate policy with a base learning rate of 0.01 and power $= 0.9$. We use random flips and rotations as data augmentation.

Training loss. Since the global scale of a scene is a fundamental ambiguity in depth estimation [13], we train the network using the scale-invariant loss in log space [42]:

$$\ell = \sqrt{\frac{1}{n}\sum_i {d_i}^2 - \frac{\lambda}{n^2}\left(\sum_i d_i\right)^2}, \qquad (4)$$

where d_i is the difference between the predicted and ground truth (log-)depth. We use $\lambda = 0.85$.

Evaluation metrics. We evaluate performance on typical depth estimation metrics [42]: absolute relative error, RMSE, log-RMSE, squared relative error, and accuracy under thresh-

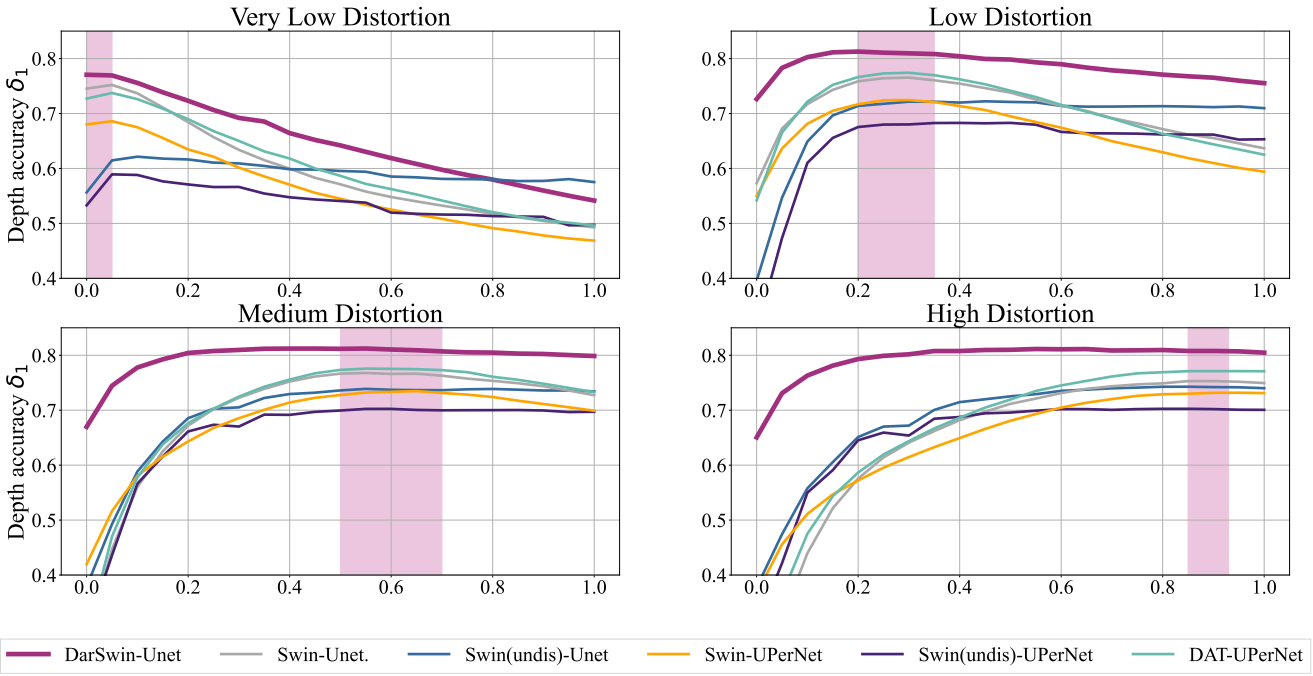


Figure 9. Depth estimation accuracy δ_1 (higher is better) as a function of test distortion for: DarSwin-Unet, Swin-Unet [6], Swin(undis)-Unet, Swin-UPerNet [32, 48], Swin(undis)-UPerNet and DAT-UPerNet [47, 48]. All methods are trained on a restricted set of lens distortion curves (indicated by the pink shaded regions): (a) Very low, (b) low, (c) medium, and (d) high distortion. We study the generalization abilities of each model by testing across $\xi \in [0, 1]$. We can see that the performance of all the baseline decreases as we move away from training distortion, but the curve for DarSwin-Unet remains relatively flat, indicating that DarSwin-Unet can generalize on unseen wide angle distortion at test time.

old (δ_i, $i \in \{1, 2, 3\}$). The paper reports results on

$$\delta_1 = \frac{1}{|D|} |\{d \in D | \max(\frac{d^*}{d}, \frac{d}{d^*}) \leq 1.25\}, \qquad (5)$$

where D, d^*, and d are the set of valid, ground truth and predicted depths, respectively. Please consult the supplementary material for results on other metrics.

5.4. Zero-shot generalization

We perform a similar generalization test as [2], we train all the baselines and DarSwin-Unet on all four training sets with different levels of distortion independently (represented by the pink shaded region in fig. 9, and evaluate them on all of the 20 test sets, as explained above. Our primary focus is on the efficacy of the network on unseen lens distortion (outside the pink shaded region). As shown in fig. 9, we can see that for all the methods including DarSwin-Unet the depth estimation accuracy δ_1 metric is highest in the pink shaded region for each training set since the model has seen those lens distortion while training. But as we move away from the pink region, the performance for the baselines decreases rapidly, as these lens distortions are not present during test time, but DarSwin-Unet maintains its performance even outside the training distortion region for ("low", "medium" and "high" distortion training sets). When DarSwin-Unet is trained in "very low" distortion, we see a decrease in performance as we move away from training distortion, but still DarSwin-Unet outperforms all the baselines.

DarSwin-Unet demonstrates better generalization capabilities across different lenses by embedding the distortion parameter within the network, as introduced by DarSwin [2]. The change in patch size, as depicted in fig. 1, allows each patch in the attention layer to be weighted based on the specific lens distortion.

For a fair comparison, the baselines Swin-Unet(undis) and Swin-UPerNet(undis) are equiped with the distortion parameter knowledge as well. However, despite this inclusion, these baselines fail to generalize effectively to other lenses. The primary reason for this is the presence of artifacts resulting from the undistortion process.

5.5. Ablations

k-NN analysis. We study the impact of the number of samples per patch on the reconstruction quality. As explained in sec. 4.1, 25 points are sampled along the radius, and 4 points are sampled along the azimuth, giving 25×4 samples for each patch. Since the k-NN layer is not trainable, ablations on this part of the model are made based on the

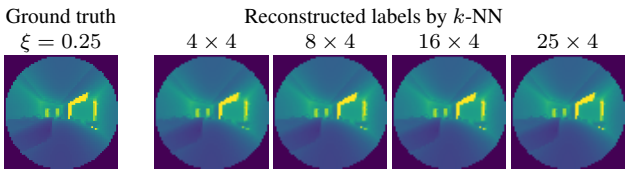


Figure 10. Visualizations of reconstructed labels by the k-NN layer considering different numbers of samples per patch. For a low number of samples (4×4 per patch), we notice some artifacts, particularly in the center. These artifacts progressively disappear when increasing the number of samples, leading to faithful reconstructions with the k-NN 25×4.

ground truth labels. We distort depth maps from Matterport3D [8] with $\xi = 0.25$. The ground truth label is sampled and reconstructed using the sample locations and the k-NN layer. To evaluate the efficiency of this layer, we calculate the Mean Absolute Error (MAE) over valid pixels between each ground truth label and its corresponding reconstructed label, and then we average values over all images. We ablate on the number of samples per patch as shown in tab. 1 and illustrated in fig. 10. We show that 25×4 samples per patch results in efficient projection from polar features to cartesian features with an error of 0.8% and 10.3ms/image compared to 3.1ms/image for 4×4 samples.

Table 1. Ablation study on the efficiency of the k-NN layer: for a limited number of samples per patch (2×2 and 4×4), we have an important error of 21.3% and 9.3%, respectively. The error reduced significantly for a number of samples equal or higher to 8×8.

# samples/patch	4×4	8×4	16×4	25×4
MAE	4.09%	2.36%	1.28%	0.8%

Sampling function. We compared the choice of sampling strategy (sec. 4.2), using depth accuracy for the model trained on each training set and tested on all $\xi \in [0, 1]$. We experiment on the three distortion curves radius vs $(\tan\theta, \theta, g(\theta))$. The generalization performance of curves with respect to $\tan(\theta)$ and θ surpasses our method $g(\theta)$ in specific cases—"very low" and "high" distortion (see fig. 11). This behavior aligns with fig. 4, as $\tan(\theta)$ benefits from dense sampling near $\xi = 0$, making it effective at low distortion but less so elsewhere. Similarly, θ performs well at high distortion due to dense sampling near $\xi = 1$. However, our proposed sampling method with respect to $g(\theta)$ consistently outperforms across all distortion levels.

6. Discussion

This paper introduces DarSwin-Unet, a novel radial-based distortion-aware encoder-decoder transformer built upon

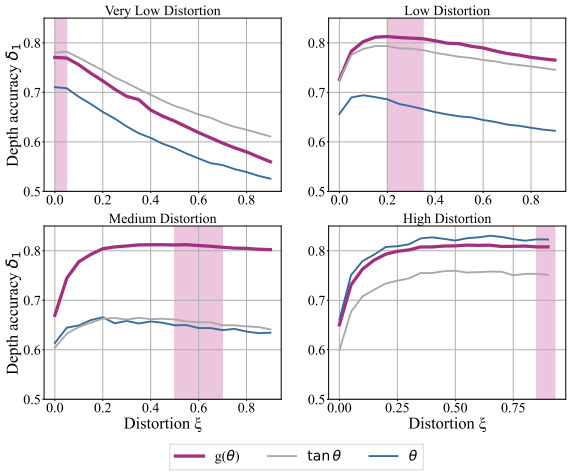


Figure 11. Depth accuracy when the model is trained across all four distortion levels—"very low," "low," "medium," and "high"—using different sampling strategies. Our proposed sampling strategy, $\mathcal{P}(g(\theta))$, demonstrates better performance in generalization across all distortion levels compared to $\mathcal{P}(\theta)$ and $\mathcal{P}(\tan\theta)$.

DarSwin [2]. DarSwin-Unet dynamically adapts its structure according to the lens distortion profile of a calibrated lens, enabling it to achieve state-of-the-art performance in zero-shot adaptation on various lenses for the depth estimation.

A key contribution of our work is the development of a novel sampling function designed to address the sparsity issues inherent in distortion-based sampling techniques introduced by [2]. This improvement is particularly important for pixel-level tasks, where sparsity in sampling has a more pronounced impact and can significantly degrade performance, unlike in classification tasks.

Limitations and future work While DarSwin-Unet shows significant advancements in distortion-aware pixel-level tasks like depth estimation, it has limitations. Scaling to high-resolution images remains challenging, but incorporating ideas from SwinV2 [31], which excels at scaling Swin Transformer architectures, could enhance its ability to handle higher resolutions. Additionally, DarSwin-Unet relies on prior knowledge of lens distortion profiles, limiting its use in uncalibrated scenarios. Future work could address this by integrating a secondary network to predict distortion parameters or developing an end-to-end model that learns distortion directly from input images.

Acknowledgments This research was supported by NSERC grant ALLRP-567654, Thales, an NSERC USRA to E. Bergeron, Mitacs and the Digital Research Alliance Canada. We thank Yohan Poirier-Ginter, Frédéric Fortier-Chouinard and Justine Giroux for proofreading.

References

[1] Ola Ahmad and Freddy Lecue. FisheyeHDK: Hyperbolic deformable kernel learning for ultra-wide field-of-view image recognition. In *Assoc. Adv. of Art. Int.*, 2022. 1, 2

[2] Akshaya Athwale, Arman Afrasiyabi, Justin Lagüe, Ichrak Shili, Ola Ahmad, and Jean-François Lalonde. Darswin: Distortion aware radial swin transformer. In *Int. Conf. Comput. Vis.*, 2023. 1, 2, 3, 4, 6, 7, 8

[3] João P. Barreto. A unifying geometric representation for central projection systems. *Comp. Vis. Img. Underst.*, 2006. 3, 5

[4] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. 2

[5] Pierre-Andre Brousseau and Sebastien Roy. Calibration of axial fisheye cameras through generic virtual central models. In *Int. Conf. Comput. Vis.*, 2019. 1

[6] Hu Cao, Yueyue Wang, Joy Chen, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, and Manning Wang. Swin-unet: Unet-like pure transformer for medical image segmentation. In *Eur. Conf. Comput. Vis. Worksh.*, 2022. 2, 4, 6, 7

[7] Oscar Carlsson, Jan E. Gerken, Hampus Linander, Heiner Spieß, Fredrik Ohlsson, Christoffer Petersson, and Daniel Persson. Heal-swin: A vision transformer on the sphere. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2024. 1, 2

[8] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *Int. Conf. on 3D Vision (3DV)*, 2017. 5, 8

[9] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Int. Conf. Comput. Vis.*, 2017. 1, 2

[10] Liuyuan Deng, Ming Yang, Hao Li, Tianyi Li, hu Bing, and Chunxiang Wang. Restricted deformable convolution-based road scene semantic segmentation using surround view cameras. *IEEE Trans. Int. Trans. Syst.*, 08 2019. 1, 2

[11] Frédéric Devernay and Olivier Faugeras. Straight lines have to be straight automatic calibration and removal of distortion from scenes of structured environments. *Mach. Vis. App.*, 13, 08 2001. 2

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Int. Conf. Learn. Represent.*, 2020. 2

[13] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Adv. Neural Inform. Process. Syst.*, 2014. 6

[14] Hao Feng, Wendi Wang, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Simfir: A simple framework for fisheye image rectification with self-supervised representation learning. In *Int. Conf. Comput. Vis.*, 2023. 2

[15] Juho Kannala and Sami S. Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(8):1335–1340, 2006. 1

[16] Byunghyun Kim, Dohyun Lee, Kyeongyuk Min, Jongwha Chong, and Inwhee Joe. Global convolutional neural networks with self-attention for fisheye image rectification. *IEEE Access*, 10:129580–129587, 2022. 2

[17] Hyungtae Kim, Eunjung Chae, Gwanghyun Jo, and Joonki Paik. Fisheye lens-based surveillance camera for wide field-of-view monitoring. In *IEEE Int. Conf. Cons. Elec.*, 2015. 1

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Adv. Neural Inform. Process. Syst.*, 2012. 2

[19] Varun Ravi Kumar. Surround-view cameras based holistic visual perception for automated driving. In *Int. Conf. on Robotics and Automation*, 202q. 2

[20] Varun Ravi Kumar, Sandesh Athni Hiremath, Stefan Milz, Christian Witt, Clement Pinnard, Senthil Yogamani, and Patrick Mader. Fisheyedistancenet: Self-supervised scale-aware distance estimation using monocular fisheye camera for autonomous driving. In *Int. Conf. on Robotics and Automation*, 2020. 2

[21] Varun Ravi Kumar, Marvin Klingner, Senthil Yogamani, Markus Bach, Stefan Milz, Tim Fingscheidt, and Patrick Mäder. Svdistnet: Self-supervised near-field distance estimation on surround view fisheye cameras. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10252–10261, 2021. 2

[22] 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 *Winter Conf. App. Comput. Vis.*, 2021. 2

[23] Varun Ravi Kumar, Senthil Yogamani, Markus Bach, Christian Witt, Stefan Milz, and Patrick Mader. Unrectdepthnet: Self-supervised monocular depth estimation using a generic framework for handling common camera distortion models. In *Int. Conf. Intell. Robots Systems*, 2023. 2

[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. In *Int. Conf. on Robotics and Automation*. IEEE, 2021. 2

[25] 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. *IEEE Robotics and Automation Letters*, 6(2):2830–2837, 2021. 2

[26] Jongsung Lee, Gyeongsu Cho, Jeongin Park, Kyongjun Kim, Seongoh Lee, Jung-Hee Kim, Seong-Gyun Jeong, and Kyungdon Joo. Slabins: Fisheye depth estimation using slanted bins on road environments. In *Int. Conf. Comput. Vis.*, 2023. 2

[27] Kang Liao, Chunyu Lin, and Yao Zhao. A deep ordinal distortion estimation approach for distortion rectification. *IEEE Trans. Image Process.*, 30:3362–3375, 2021. 2

[28] Kang Liao, Chunyu Lin, Yao Zhao, and Mai Xu. Model-free distortion rectification framework bridged by distortion

distribution map. *IEEE Trans. Image Process.*, 29:3707–3718, 2020. 2

[29] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2022. 2

[30] Shuying Liu and Weihong Deng. Very deep convolutional neural network based image classification using small training sample size. In *IAPR Asian Conf. on Pattern Recog.*, 2015. 2

[31] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin transformer v2: Scaling up capacity and resolution, 2022. 8

[32] 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 *Int. Conf. Comput. Vis.*, 2021. 2, 3, 6, 7

[33] Christopher Mei and Patrick Rives. Single view point omnidirectional camera calibration from planar grids. In *Int. Conf. on Robotics and Automation*, 2007. 3, 5

[34] R. Melo, M. Antunes, J. P. Barreto, G. Falcão, and N. Gonçalves. Unsupervised intrinsic calibration from a single frame using a "plumb-line" approach. In *Int. Conf. Comput. Vis.*, 2013. 1

[35] Elad Plaut, Erez Ben Yaacov, and Bat El Shlomo. 3d object detection from a single fisheye image without a single fisheye training image. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2021. 1, 2

[36] Clément Playout, Ola Ahmad, Freddy Lécué, and Farida Cheriet. Adaptable deformable convolutions for semantic segmentation of fisheye images in autonomous driving systems. *CoRR*, abs/2102.10191, 2021. 1, 2

[37] S. Ramalingam and Peter Sturm. A unifying model for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(7):1309–1319, 2017. 1

[38] Hazem Rashed, Eslam Mohamed, Ganesh Sistu, Varun Ravi Kumar, Ciaran Eising, Ahmad El-Sallab, and Senthil Yogamani. Generalized object detection on fisheye cameras for autonomous driving: Dataset, representations and baseline. In *Winter Conf. App. Comput. Vis.*, 2021. 2

[39] Hazem Rashed, Eslam Mohamed, Ganesh Sistu, Varun Ravi Kumar, Ciaran Eising, Ahmad Sallab, and Senthil Yogamani. Fisheyeyolo: Object detection on fisheye cameras for autonomous driving. In *Adv. Neural Inform. Process. Syst.*, 12 2020. 2

[40] Dieter Schmalstieg and Tobias Höllerer. Augmented reality: Principles and practice. In *IEEE Virt. Reality*, 2017. 1

[41] Hao Shi, Yu Li, Kailun Yang, Jiaming Zhang, Kunyu Peng, Alina Roitberg, Yaozu Ye, Huajian Ni, Kaiwei Wang, and Rainer Stiefelhagen. Fishdreamer: Towards fisheye semantic completion via unified image outpainting and segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2023. 2

[42] Chang Shu, Ziming Chen, Lei Chen, Kuan Ma, Minghui Wang, and Haibing Ren. Sidert: A real-time pure transformer architecture for single image depth estimation. *arXiv preprint arXiv:2204.13892*, 2022. 6

[43] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott E. Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Feature pyramid networks for object detection. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 2

[44] Álvaro Sáez, Luis M. Bergasa, Eduardo Romeral, Elena López, Rafael Barea, and Rafael Sanz. Cnn-based fisheye image real-time semantic segmentation. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, 2018. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inform. Process. Syst.*, 2017. 2

[46] Xuan Wei, Zhidan Ran, and Xiaobo Lu. Dcpb: Deformable convolution based on the poincare ball for top-view fisheye cameras. In *Int. Conf. Comput. Vis.*, 2023. 2

[47] Zhuofan Xia, Xuran Pan, Shiji Song, Li Erran Li, and Gao Huang. Vision transformer with deformable attention. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2, 6, 7

[48] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Eur. Conf. Comput. Vis.*, 2018. 6, 7

[49] Yuwen Xiong, Renjie Liao, Hengshuang Zhao, Rui Hu, Min Bai, Ersin Yumer, and Raquel Urtasun. Upsnet: A unified panoptic segmentation network. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 2

[50] Z. Xue, N. Xue, G. Xia, and W. Shen. Learning to calibrate straight lines for fisheye image rectification. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2

[51] Qingan Yan, Pan Ji, Nitin Bansal, Yuxin Ma, Yuan Tian, and Yi Xu. Fisheyedistill: Self-supervised monocular depth estimation with ordinal distillation for fisheye cameras. In *Winter Conf. App. Comput. Vis.*, 2021. 2

[52] Shangrong Yang, Chunyu Lin, Kang Liao, Chunjie Zhang, and Yao Zhao. Progressively complementary network for fisheye image rectification using appearance flow. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021. 2

[53] 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. 2

[54] Shangrong Yang, Chunyu Lin, Kang Liao, and Yao Zhao. Dual diffusion architecture for fisheye image rectification: Synthetic-to-real generalization. In *Int. Conf. Comput. Vis.*, 2023. 2

[55] Yaozu Ye, Kailun Yang, Kaite Xiang, Juan Wang, and Kaiwei Wang. Universal semantic segmentation for fisheye urban driving images. *IEEE Int. Conf. Syst. Man Cyber.*, pages 648–655, 2020. 2

[56] Xiaoqing Yin, Xinchao Wang, Jun Yu, Maojun Zhang, Pascal Fua, and Dacheng Tao. Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification. In *Eur. Conf. Comput. Vis.*, 2018. 1, 2

[57] Xianghua Ying and Zhanyi Hu. Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model. In *Eur. Conf. Comput. Vis.*, 2004. 3

[58] Senthil Yogamani, Ciaran Hughes, Jonathan Horgan, Ganesh Sistu, Sumanth Chennupati, Michal Uricar, Stefan Milz, Martin Simon, Karl Amende, Christian Witt, Hazem Rashed,

Sanjaya Nayak, Saquib Mansoor, Padraig Varley, Xavier Perrotton, Derek Odea, and Patrick Pérez. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Int. Conf. Comput. Vis.*, 2019. 1, 2, 5, 6

[59] Jiaming Zhang, Kailun Yang, Hao Shi, Simon Reiß, Kunyu Peng, Chaoxiang Ma, Haodong Fu, Philip H. S. Torr, Kaiwei Wang, and Rainer Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2022. 2

[60] Mi Zhang, Jian Yao, Menghan Xia, Kai Li, Yi Zhang, and Yaping Liu. Line-based multi-label energy optimization for fisheye image rectification and calibration. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015. 1, 2

[61] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019. 1, 2