ACDNet: Adaptively Combined Dilated Convolution for Monocular Panorama Depth Estimation

Chuanqing Zhuang[1*], **Zhengda Lu**[1*], **Yiqun Wang**[2,3], **Jun Xiao**[1†], **Ying Wang**[1]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] College of Computer Science, Chongqing University
[3] KAUST

zhuangchuanqing19@mails.ucas.ac.cn, {luzhengda,xiaojun,ywang}@ucas.ac.cn, yiqun.wang@cqu.edu.cn

Abstract

Depth estimation is a crucial step for 3D reconstruction with panorama images in recent years. Panorama images maintain the complete spatial information but introduce distortion with equirectangular projection. In this paper, we propose an ACDNet based on the adaptively combined dilated convolution to predict the dense depth map for a monocular panoramic image. Specifically, we combine the convolution kernels with different dilations to extend the receptive field in the equirectangular projection. Meanwhile, we introduce an adaptive channel-wise fusion module to summarize the feature maps and get diverse attention areas in the receptive field along the channels. Due to the utilization of channel-wise attention in constructing the adaptive channel-wise fusion module, the network can capture and leverage the cross-channel contextual information efficiently. Finally, we conduct depth estimation experiments on three datasets (both virtual and real-world) and the experimental results demonstrate that our proposed ACDNet substantially outperforms the current state-of-the-art (SOTA) methods. Our codes and model parameters are accessed in *https://github.com/zcq15/ACDNet*.

Introduction

The panoramic camera is a new type of camera to capture images with $180° \times 360°$ field of view (FoV), which is convenient to obtain omnidirectional spatial information in a single shot without the post-calibration and stitching. With its wide usage in the fields such as virtual reality (VR) and security monitoring in recent years, panorama depth estimation is a crucial step in a variety of downstream applications, such as semantic segmentation, layout recovery, and 3D reconstruction, to name a few.

Generally, panorama images are represented as images on the sphere grid for warp and weft by the equirectangular projection (ERP). However, the geometric structure in the higher latitude areas is distorted since the spatial sampling rate changes with latitude. Therefore, accurate depth estimation is difficult with conventional convolution networks in these areas.

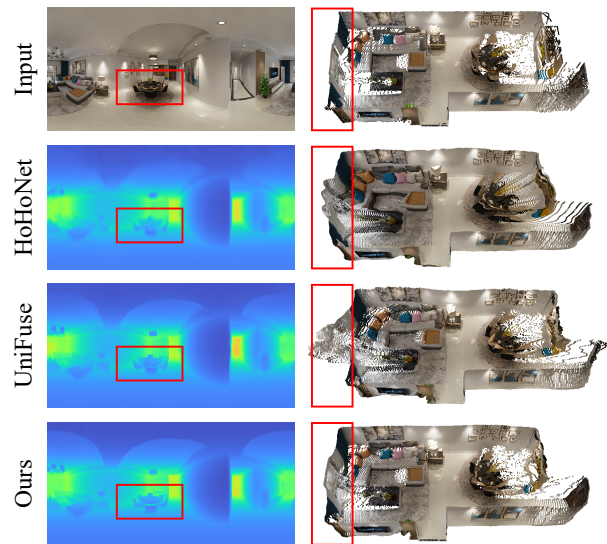*Joint first authors with equal contribution

†Corresponding author

Figure 1: Reconstructed results of different models. Left: the panorama color image and the predicted depth maps. Right: point clouds generated from the ground-truth and predicted depth maps with inverse equirectangular projection. Note that our predicted depth map has more clear edges and also generates a more accurate point cloud in both overall shape and object details.

Early works (Cohen et al. 2017, 2018) define the spherical CNNs to process the spherical signals but cause the high resource expenditure. And some others (Su and Grauman 2017; Tateno, Navab, and Tombari 2018; Coors, Condurache, and Geiger 2018; Fernandez-Labrador et al. 2020) propose the different custom convolutions to deform the convolution kernels according to the geometric structures in the 3D space of the ERP coordinates. These adaptive convolution kernels expand the receptive fields near the poles according to the corresponding latitude coordinates. However, these methods still have a great potential to be further improved.

On the other hand, BiFuse (Wang et al. 2020a) and Uni-Fuse (Jiang et al. 2021) project the ERP image to the cube-

map images to solve the distortion with the perspective projection. However, due to the limitation of the FoV in the cubemap branch, the overall layout in the reconstructed scenes can not be well restored. Besides, HoHoNet (Sun, Sun, and Chen 2021) and SliceNet (Pintore et al. 2021) extract the horizontal 1D feature maps from gravity-aligned equirectangular projection and recover the dense 2D predictions. However, it is hard to recover the details in the columns from 1D features (see Fig.1). Thus, the balance between quantitative results and visual effects still needs to be considered.

In this paper, we propose the ACDNet based on the adaptively combined dilated convolution for the panoramic monocular depth estimation. We combine the convolution kernels with different dilations to extend the receptive field in the equirectangular projection. Meanwhile, we use an adaptive channel-wise fusion module to summarize the feature maps and get diverse attention areas in the receptive field along different channels. Different from methods (Su and Grauman 2017; Tateno, Navab, and Tombari 2018; Coors, Condurache, and Geiger 2018; Fernandez-Labrador et al. 2020) that calculate the shapes of convolution kernels according to the latitude coordinates, we learn the focused areas in different feature channels that help the network to capture the cross-channel contextual information. Finally, we evaluate our method on both virtual and real-world panoramic RGB-D datasets. The experimental results show that our ACDNet and the adaptively combined dilated convolution outperform the current state-of-the-art methods.

In summary, the main contributions of this work can be summarized as follows:

1. We propose the adaptively combined dilated convolution to process the panorama images for monocular depth estimation, and it can be easily embedded into convolution networks by replacing the regular convolution.

2. The interest areas can be obtained by learning different attention scores in different channels, which is more suitable for panoramic images than explicitly deforming convolution kernels in different latitudes.

3. We perform the monocular panorama depth estimation experiments on both virtual and real-world RGB-D panorama datasets, which outperforms the SOTA methods in both quantitative metrics and visual effects.

Related Work

In this section, we describe the overview of researches on panorama depth estimation and simply introduce the applications of dilated convolution in CNNs.

Panorama Depth Estimation

Depth estimation is an important step for 3D reconstruction, and panorama images can capture the omnidirectional spatial information for the global structure, which conduces to recover the depth in areas with weak textures.

OmniDepth (Zioulis et al. 2018) first proposes the RectNet to estimate the depth map with a single panorama image and shows better performance than individually processing the different views of cubemap projection (CMP). However, this method is limited by the distortion of the geometric structures and the decrease of the FoV near the poles for panoramic images in the equirectangular projection. There are some main types of existing methods to solve this problem.

Firstly, some methods (Coors, Condurache, and Geiger 2018; Tateno, Navab, and Tombari 2018; Fernandez-Labrador et al. 2020; Eder et al. 2019) deform the convolution kernels to adaptively extend the receptive fields of custom convolutions. Specifically, SphereNet (Coors, Condurache, and Geiger 2018) and DistConv (Tateno, Navab, and Tombari 2018) calculate the sampling positions for the convolution kernels with inverse gnomonic projection, and CFL (Fernandez-Labrador et al. 2020) defines the convolution over the field of view on the spherical surface with longitudinal and latitudinal angles. Besides, mapped convolution (Eder et al. 2019) proposes a more general method to process images of any structured representation by accepting the corresponding mapping function. These methods all produce the different convolution kernels with the change of latitude coordinates in the ERP.

Secondly, some other methods introduce the additional CMP branch with perspective projection into the network. BiFuse uses their proposed bi-projection fusion module to fuse the feature maps in two complete encoder-decoder branches. Furthermore, UniFuse removes the CMP decoder branch and proposes a more effective unidirectional fusion module. However, due to the limitation of the FoV, the CMP branch can not extract good features from areas with weak textures, e.g., the ceilings and the floors, which cripples the ability for fused features to express the spatial structures. Last, recent works HoHoNet and SliceNet compress the 2D features to 1D features from gravity-aligned panoramic images, then they apply the RNNs to capture the global context information.

There are also still other methods to estimate panoramic depth with different strategies, including deformable convolution kernels (Cheng et al. 2020; Chen et al. 2021), geometric guidance (Eder, Moulon, and Guan 2019; Jin et al. 2020; Zeng, Karaoglu, and Gevers 2020), self-supervised/unsupervised learning (Zioulis et al. 2019; Zhou, Wang, and Yang 2020), and stereo matching (Wang et al. 2020b). These existing methods have achieved good results on depth estimation with panoramic cameras. But there is still much room for improvement in terms of quantitative results or visual effects.

Dilated Convolution

It is proved that dilated convolution is an effective tool for the increment of the receptive field without additional parameters and down-sampling. Apart from semantic segmentation (Chen et al. 2018) and object detection (Liu et al. 2016), dilated convolution is also widely utilized on some other tasks such as depth estimation.

Earlier works (Yu, Koltun, and Funkhouser 2017; Ma et al. 2018) simply embed the dilated convolutions in the network. And many works use the atrous spatial pyramid pooling (ASPP) (Chen et al. 2018) or a similar module to
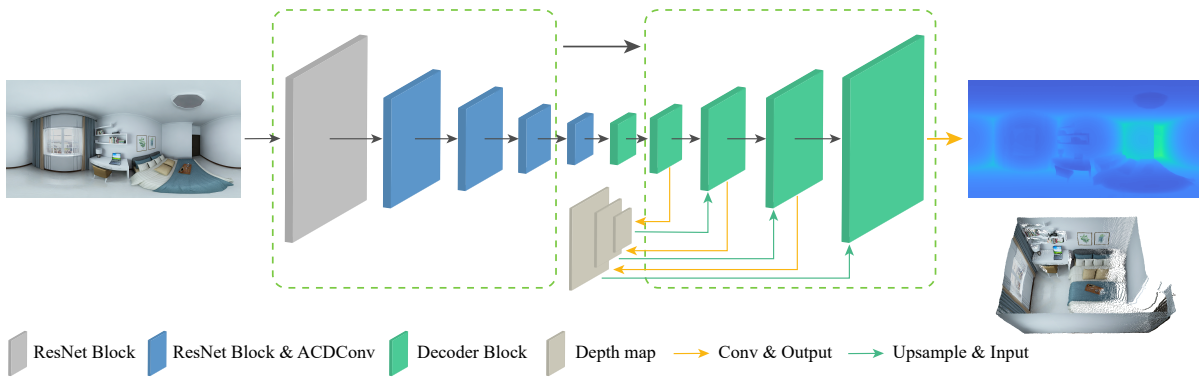
Figure 2: The architecture of our ACDNet.

model contextual information. Some works (Fu et al. 2018; Fang et al. 2020; Zhang et al. 2020; Lee et al. 2021) introduce ASPP to aggregate multi-scale contextual information for better monocular depth estimation. Besides, MSDC-Net (Tian et al. 2019) combines the res-block module with different dilated rates to build the irregular shape ResNet (He et al. 2016) module. CrossGuidance (Lee et al. 2020) proposes a residual atrous spatial pyramid (RASP) block to analyze the large input images. And MAPUnet (Yang et al. 2021) develops the multi-layers DenseASPP with more scales to cover more pixels.

Different from modules similar to ASPP, we combine the dilated convolutions as an equivalent large kernel convolution and apply it to replace the regular convolution layers in ResNet blocks. This operator enlarges the receptive field and produces a variety of interest areas in the receptive field.

Approach

For the reconstruction of the indoor scene with a single panoramic image, we propose the ACDNet with adaptively combined dilated convolution (ACDConv) layers to estimate the depth map. In the following text, we first present the architecture of the ACDNet, then we show the implementation of the ACDConv to extract feature maps from panorama images. Finally, we introduce the loss function in our approach.

Architecture

We propose the ACDNet to estimate the depth map with a single panorama image as illustrated in Fig.2. In general, the ACDNet is a conventional network with ResNet blocks based on the ACDConv and the iterative depth prediction process. Specifically, given an input panoramic color image, the encoder extracts feature maps in five downsampling scales with the ResNet blocks. Here, the 3×3 convolution layers are replaced by our proposed ACDConv, the detail is introduced in the following sub-section. Second, the decoder upsamples the feature maps with up-convolution modules (Laina et al. 2016) and produces the depth maps in 4 different scales. The first coarse depth map D_0 is generated at the 1/8 downsampling level (level 0), then the subsequent residual maps $\{R_i\}_{i=1}^{3}$ are produced in the following decoder blocks. The depth map D_i at level i is formulated as

(a) Padding around poles
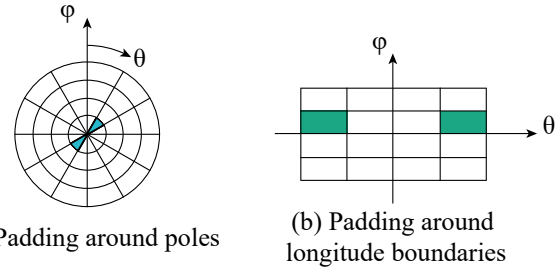
(b) Padding around longitude boundaries

Figure 3: The circular padding for panoramic feature maps. The elements with the same color are selected to pad each other.

$D_i = \tilde{D}_{i-1} + R_i$, where $\tilde{D}_{i-1}$ is up-sampled from D_{i-1} with bilinear interpolation.

Besides, the circular padding (Wang et al. 2018) is utilized to maintain a complete and continuous spatial field of view for panoramic images. As shown in Fig.3, we select feature items along the longitudinal direction near the poles and the latitudinal direction near the other boundaries. The circular padding maintains the continuity of spatial information on the sphere surface and avoids the invalid padding elements for dilated convolutions.

Adaptively Combined Dilated Convolution

The panorama image expands the spherical imaging result to a rectangular image, which causes the narrow FoV near the poles for regular convolutions. Previous works have developed different custom convolutions to extend FoV near the poles. In this paper, we combine the regular convolutions with different dilations to increase the FoV. Moreover, we introduce an adaptive channel-wise fusion (ACF) module to aggregate the feature maps and get diverse attention areas in the receptive field along the channels.

The details of our proposed ACDConv are illustrated in Fig.4. First, given the input features $F_{input} \in \mathbb{R}^{C \times H \times W}$, we use the different convolutions with a group of dilation settings to extract feature maps $\{F^i\}_{i=1}^{4} \in \mathbb{R}^{C' \times H \times W}$ from the input features in parallel. Then, a learnable ACF module is applied to integrate the feature maps. Specifically,
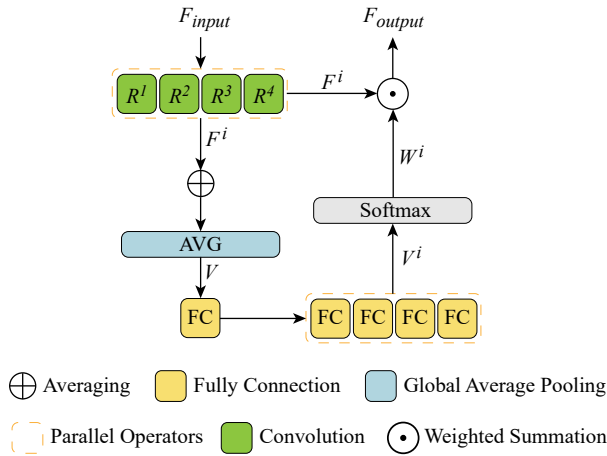
Figure 4: Our adaptively combined dilated convolution layer. The label R^n in the figure means the n-th choice of the four dilation rate settings.

we first get the intermediate mean feature and utilize the global average pooling to obtain a vector $V \in \mathbb{R}^{C'}$. After that, the fully connected layers predict the probability vectors $\{V^i\}_{i=1}^4 \in \mathbb{R}^{C'}$ for different feature maps, and the softmax function is applied to produce the channel-wise fusion weights as follows:

$$W_c^i = \frac{\exp V_c^i}{\sum_{j=1}^4 \exp V_c^j} \qquad (1)$$

Finally, the feature maps F^i from different convolutions are summarized with the channel-wise weights W^i to generate the final feature maps $F_{output} = \sum_i F^i W^i$ and get a large receptive field.

On the other hand, we draw the receptive fields of different convolutions in Fig.5, including the regular 3×3 convolution, the custom convolution with inverse gnomonic projection in SphereNet, and our adaptively combined dilated convolution. As shown in Fig.5 (a), the vanilla convolution always keeps the 3×3 receptive field in the different latitudes of the ERP, and SphereNet deforms the convolution kernel according to the latitudes to extend the receptive field, especially in the poles areas (see Fig.5 (b)). Our combined convolution keeps the shape of a large receptive field in the different latitudes of the ERP as shown in Fig.5 (c). More importantly, different attention weights in diverse areas of the combined receptive field along the channels can be acquired after the weighted summarization. Besides, we set the dilation settings as 1×1, 1×2, 1×4, and 2×1 in our experiments.

Loss Function

In our approach, we use the BerHu (Laina et al. 2016) loss function to supervise the training process for the network, which is formulated as:

$$L(d_i, \widetilde{d}_i) = \begin{cases} |\Delta| & |\Delta| \leq c \\ \frac{|\Delta|^2 + c^2}{2c} & |\Delta| > c \end{cases} \qquad (2)$$

where $\Delta = d_i - \widetilde{d}_i$, and $d_i, \widetilde{d}_i$ are the estimated depth and the ground truth on pixel i of input image respectively.

For each input image, the parameter c is set as

$$c = \frac{1}{5} \max_i |d_i - \widetilde{d}_i| \qquad (3)$$

Finally, we apply the BerHu loss on D_3, the shape of which is the same as the input image, and $\{D_i\}_{i=0}^2$ are part of the components of D_3.

Experiments

In this section, we first introduce our experiments, including datasets, implementation details, and evaluation metrics. Then, we provide the qualitative and quantitative comparisons of our network with state-of-the-art approaches. Finally, we perform the ablation experiments to validate the effectiveness of our network structure. All experiments were conducted on a server computer equipped with an Intel(R) Xeon(R) Gold 6130 CPU processor, 256GB of RAM, and an NVIDIA TITAN RTX 24GB graphics card.

Implementation

Datasets We carry out experiments on both virtual and real-world datasets, including Stanford2D3D (Armeni et al. 2017), Matterport3D (Chang et al. 2017), and Structured3D (Zheng et al. 2020). Both Stanford2D3D and Matterport3D are scanned with RGB-D cameras in the real-world scenes, and they include $1,413$ and $10,800$ RGB-D panoramic views respectively. While Structured3D is rendered with synthetic scenes, and it contains over $196k$ RGB-D panorama images. For Stanford2D3D and Matterport3D, we follow their official splits with entire panoramic RGB-D pairs to train and test the network. For Structured3D, we just utilize the subset with *rawlight* illumination and *full* furniture settings. The subset includes $21,835$ panoramic RGB-D image pairs, and we follow the official scene split for training and testing. Moreover, we follow the process strategy for Matterport3D as previous works to merge the 18-views perspective depth images and the rendered skybox color images to panoramic RGB-D image pairs.

Implementation Details We implement our network on the PyTorch (Paszke et al. 2019) platform. We train our network for 100 epochs on Stanford2D3D, 60 epochs on Matterport3D, and 60 epochs on Structured3D with Adam (Kingma and Ba 2015) optimizer respectively, the learning rate is set as 1e-4 in all the experiments. Meanwhile, we set the image size as 512×1024 with the batch size of 6 on an NVIDIA TITAN RTX graphics card.

Evaluation Metrics We adopt five widely-used evaluation metrics used in previous works to evaluate our method quantitatively, including mean absolute error (MAE), root mean square error (RMSE), logarithmic root mean square error (RMSElog), absolute relative error (Abs Rel), and threshold percentage (δ^n), which can be formulated as:

· $MAE = \frac{1}{N} \sum_{i=1}^N |d_i - \widetilde{d}_i|$;

· $RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N |d_i - \widetilde{d}_i|^2}$;

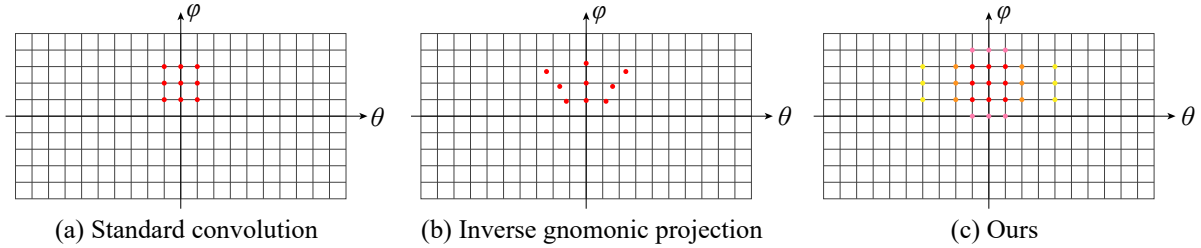(a) Standard convolution   (b) Inverse gnomonic projection   (c) Ours

Figure 5: Receptive fields (RF) of different convolutions. In sub-figure (c), the items with the same color in a sub-area of a receptive field have the same interest scores.

Dataset	Method	MAE↓	RMSE↓	RMSElog↓	AbsRel↓	$\delta^1 \uparrow$	$\delta^2 \uparrow$	$\delta^3 \uparrow$
Stanford2D3D	BiFuse	0.2343	0.4142	0.0787	0.1209	86.60	95.80	98.60
	UniFuse	0.2082	0.3691	0.0721	0.1114	87.11	96.64	98.82
	HoHoNet	0.2027	0.3834	0.0668	0.1014	**90.54**	96.93	98.86
	SliceNet	**0.1757**	0.3509	0.0801	0.0995	90.29	96.26	98.44
	SphereNet	0.2253	0.3833	0.0786	0.1234	85.39	95.67	98.33
	Ours	0.1870	**0.3410**	**0.0664**	**0.0984**	88.72	**97.04**	**98.95**
Matterport3D	BiFuse	0.3470	0.6259	0.1134	0.2048	84.52	93.19	96.32
	UniFuse	0.2814	0.4941	0.0701	0.1063	88.97	96.23	98.31
	HoHoNet	0.2862	0.5138	0.0871	0.1488	87.86	95.19	97.71
	SliceNet	0.3296	0.6133	0.1045	0.1764	87.16	94.83	97.16
	SphereNet	0.3167	0.5212	0.0778	0.1258	84.34	95.49	98.17
	Ours	**0.2670**	**0.4629**	**0.0646**	**0.1010**	**90.00**	**96.78**	**98.76**
Structured3D	BiFuse	0.0562	0.1100	0.0295	0.0401	98.19	99.41	99.72
	UniFuse	0.0617	0.1167	0.0324	0.0458	97.65	99.28	99.69
	HoHoNet	0.0549	0.1088	0.0316	0.0408	97.97	99.35	99.70
	SliceNet	0.0660	0.1290	0.0444	0.0496	97.25	99.09	99.54
	SphereNet	0.0664	0.1161	0.0368	0.0491	97.58	99.36	99.71
	Ours	**0.0454**	**0.0924**	**0.0291**	**0.0327**	**98.74**	**99.59**	**99.82**

Table 1: Quantitative comparison on different datasets. The best result of each measurement is marked in **bold** font. Here we re-train the previous works on Structured3D with their source codes at the resolution of 512×1024, and the training strategies are the same as ours.

· $RMSElog = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |\log d_i - \log \widetilde{d_i}|^2}$;

· $AbsRel = \frac{1}{N}\sum_{i=1}^{N} |d_i - \widetilde{d_i}|/\widetilde{d_i}$;

· Threshold percentage δ^n is the percentage of pixels satisfying $\max(\frac{d_i}{\widetilde{d_i}}, \frac{\widetilde{d_i}}{d_i}) < 1.25^n$.

Following the previous methods, we clip the estimated depth maps to $10m$ without scale calibration when calculating the evaluation metrics.

Comparison Experiments

In this sub-section, we provide the quantitative comparison and visual comparison to prove the effectiveness of our method.

Quantitative Comparison We compare our ACDNet with previous works on the three above-mentioned datasets, and the quantitative results are shown in Tab.1. Our ACD-Net outperforms previous works for most metrics on Stan-

ford2D3D and all metrics on Matterport3D and Structured3D. Note that the results of SliceNet on Stanford2D3D are produced by the fixed parameters in SliceNet's Github repository [1] and differ from the original values in its paper. Specifically, our results exceed the previous state-of-the-art method by 5.1% in MAE metric on Matterport3D and 17.3% on Structured3D as well as reduce the AbsRel metric by 1.1%, 5.0%, and 18.5% on the three datasets. Besides, we also compare our method with the SphereNet that uses a custom convolution, where we implement by using the same framework and replace the convolution layers in ResNet50 with that in SphereNet. According to the results in Tab.1, our ACDNet with ACDConv also outperforms the distortion-aware convolution with inverse gnomonic projection in SphereNet. On the one hand, the ACDConv expands the receptive field to get a large spatial FoV. On the other hand, it focuses on the various areas in the receptive field along different channels, which makes the convolution ker-

---

[1]https://github.com/crs4/SliceNet

nels learn a variety of latent kernel shapes in different channels, and our network can accommodate the diverse geometric relationships in ERP.

Visual Comparison   Furthermore, Fig.6 shows our visual comparison results with SOTA methods on three datasets. As shown in Fig.6, our ACDNet predicts more accurate and detailed depth maps with better visual effects. Firstly, we recover more clear and accurate walls in the invisible areas, as shown in the first row of Fig.6. It can be easily noticed that our ACF module plays a key role in capturing the spatial global context and the circular padding helps to keep the spatial continuity of the panorama images on the sphere. Compared to our single-branch, the cubemap branch proposed in the UniFuse extracts features with weak texture in the ceiling areas of a narrow FoV, which leads to bad performance in the obscured areas. Secondly, our ACDNet estimates more object details in the depth maps, such as the bookcase in the second row and the bathtub in the third row in Fig.6. Our network also performs well in distinguishing between the background and the foreground objects with similar depth values as shown in the last three rows in Fig.6. Moreover, our ACDNet generates more accurate edges in the depth maps, such as the wall in the third row and the ceiling lamps in the last row in Fig.6. These results demonstrate the better performance of our ACDNet in the depth maps estimation with monocular panorama images.

Ablation Studies

To further verify the effectiveness of our ACDNet, we introduce some groups of ablation studies on the Stanford2D3D dataset in this section. First, we conduct some ablation studies on our ACDConv, including the different parts, the different dilation directions, and the number of dilated convolutions. Then, we compare the results of different padding methods and study the advantage of iterative depth prediction. Finally, we test the network with different ResNet backbones. In addition, we also compare the model complexity and inference time with existing methods. In all of these experiments, we use the same hyper-parameters and training strategy on Stanford2D3D.

Adaptively Combined Dilated Convolution   Here, four experiments are executed to study the roles of the dilated convolution and the ACF module in our ACDConv as shown in Tab.2. First, we remove the ACDConv and use the original ResNet backbone with the regular convolution in our network as the baseline. Then we introduce the ACDConv but replace our ACF module with a simple average operator, denoted as *Simple*. Moreover, we test two other fusion strategies in ACDConv instead of our channel-wise fusion strategy. Specifically, given the intermediate feature maps $\{F^i\}_{i=1}^4 \in \mathbb{R}^{C' \times H \times W}$, the *Row-wise* strategy adds F^i and squeezes the results to row-wise feature vector $V \in \mathbb{R}^H$ with averaging operators, then the following MLPs generate the row-wise normalized probability $\{W^i\}_{i=1}^4 \in \mathbb{R}^H$ for fusion process. By contrast, the *Pixel-wise* strategy does not squeeze the feature maps and produces pixel-wise normalized probability $\{W^i\}_{i=1}^4 \in \mathbb{R}^{C' \times H \times W}$ with 1×1 convolu-

Method	MAE↓	RMSE↓	RMSElog↓	AbsRel↓
Baseline	0.2104	0.3620	0.0746	0.1148
Simple	0.2037	0.3582	0.0689	0.1075
Row-wise	0.2096	0.3694	0.0759	0.1124
Pixel-wise	0.2096	0.3659	0.0720	0.1090
Ours	**0.1870**	**0.3410**	**0.0664**	**0.0984**
ASPP	0.1990	0.3633	0.0700	0.1032

Table 2: Ablation studies about ACDConv. In the table, we use the different convolutions in the ResNet50 blocks to build the network.

Method	MAE↓	RMSE↓	RMSElog↓	AbsRel↓
Baseline	0.2104	0.3620	0.0746	0.1148
X-axis	0.1953	0.3559	0.0679	0.1028
Y-axis	0.1965	0.3540	0.0684	0.1028
Ours	**0.1870**	**0.3410**	**0.0664**	**0.0984**

Table 3: Ablation studies about different dilation directions.

tions. Finally, we also test the popular ASPP module in our baseline which combines the different dilated convolutions to capture multi-scale context.

As shown in Tab.2, simply averaging the features from different dilated convolutions or utilizing the ASPP module can improve the performance on panoramic images, which shows that the large receptive field of convolution contributes to this task. Our adaptive aggregation operator produces various attention areas on the receptive field and can adapt to the different geometric properties in the different latitudes of panoramic images, thus our full ACDConv outperforms these two methods. Besides, the adaptively combined dilated convolution with row-wise or pixel-wise fusion strategy hardly improves the results or even makes it worse. Thus, explicitly combining different dilation at different latitudes or different pixel locations is not suitable for panoramic images and our adaptively learning the receptive field in different feature channels is a more applicable solution, which also explains the reason that the custom distortion-aware convolution does not perform well in this task.

Due to the deployment of the dilation convolutions along both the x-axis and the y-axis, we study the effects of different dilation directions. Specifically, we separately test the dilation only along the x-axis or the y-axis with the combined dilation settings of 1, 2, 3, and 4. As shown in Tab.3, dilation settings along only the x-axis or y-axis can improve the MAE metric to 0.1953 and 0.1965 respectively. As the areas near the poles have narrow FoV along the latitude direction in the ERP coordinates, the dilated convolution along the latitude contributes to addressing this problem. Moreover, despite that the ERP expression keeps uniform FoV along the longitude direction, the areas near the poles mainly include the weak textures and geometric structures, e.g., the ceiling and the floor. Thus, the dilation along the y-axis introduces more spatial information in these areas in estimating accu-
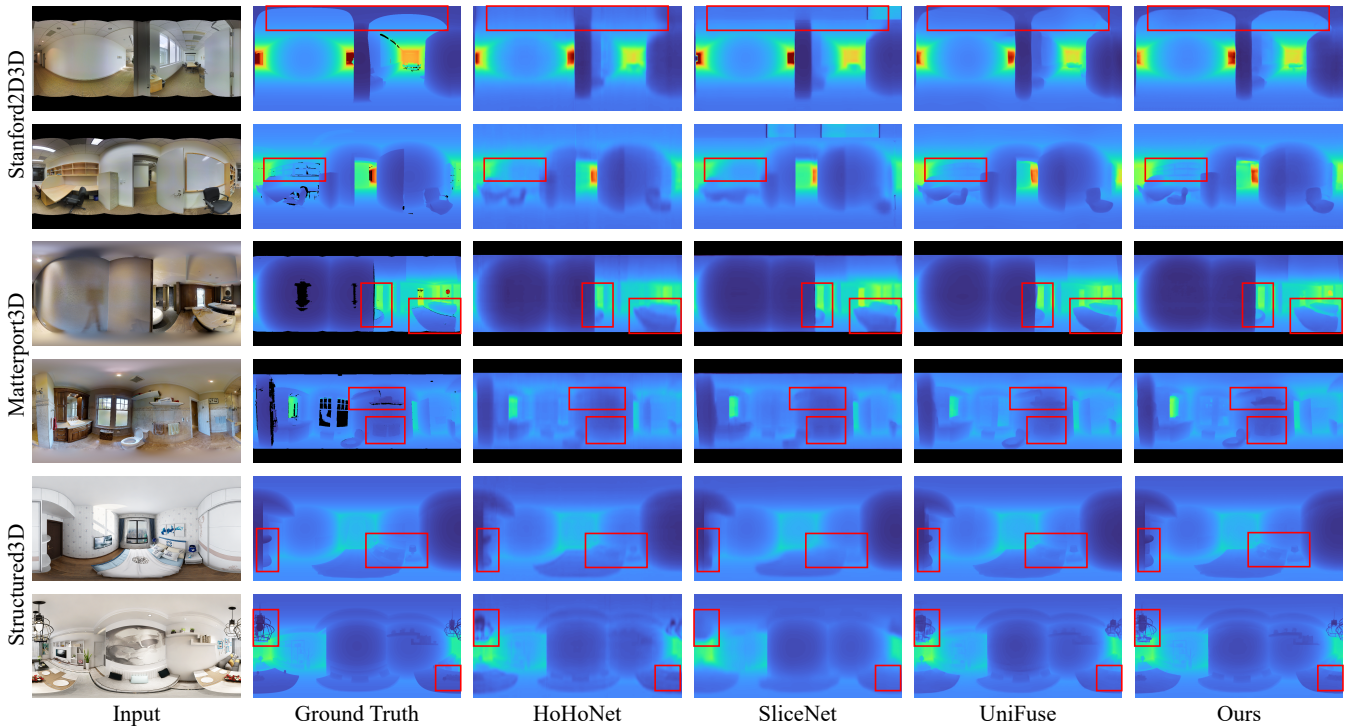
Figure 6: Depth maps comparison with other methods. The area with zero values in ground truth means the missing area of depth maps.

Dilations	MAE↓	RMSE↓	RMSElog↓	AbsRel↓
Baseline	0.2104	0.3620	0.0746	0.1148
Two	0.2038	0.3632	0.0703	0.1088
Three	0.1971	0.3573	0.0687	0.1023
Four (Ours)	**0.1870**	**0.3410**	**0.0664**	**0.0984**
Five	0.1963	0.3561	0.0689	0.1047

Table 4: Ablation studies about different dilation numbers.

Padding	MAE↓	RMSE↓	RMSElog↓	AbsRel↓
ZeroPad	0.1948	0.3526	0.0684	0.1045
LRPad	0.1935	0.3503	0.0670	0.1025
CirPad	**0.1870**	**0.3410**	**0.0664**	**0.0984**

Table 5: Ablation studies about different padding methods.

Methods	MAE↓	RMSE↓	RMSElog↓	AbsRel↓
Baseline w/ iter	0.2104	0.3620	0.0746	0.1148
Baseline w/o iter	0.2287	0.3923	0.0796	0.1200
Ours w/ iter	**0.1870**	**0.3410**	**0.0664**	**0.0984**
Ours w/o iter	0.2017	0.3650	0.0694	0.1022

Table 6: Ablation studies about iterative depth prediction.

rate depth. Therefore, our dilation settings along both directions have the best performance (see in Tab.3).

We also discuss the impact of the dilation number in the ACDConv shown in Tab.4. More dilations could bring larger receptive fields and better depth estimation performance but increase the network complexity and aggravate training difficulty at the same time. When the dilation number is 5 with an additional 1×8 dilation setting, the network is overfitted, which will make performance worse. Thus, the dilation number is set to 4 in our experiments.

Padding Method   In the proposed ACDNet, we apply circular padding to get continuous features on the sphere. In this sub-section, we also test the effects of zero padding and left-right padding as shown in Tab.5. We observe that the results have gradual improvements by zero padding, left-right padding, and circular padding. The root cause is that proper padding avoids introducing abundant invalid elements into the dilated convolutions in the boundary regions. Mean-

while, this also demonstrates the effectiveness of the complete and continuous spatial information for depth estimation in the panoramic images.

Iterative Depth Prediction   In this sub-section, we separately test the role of iterative depth prediction and our ACDConv. According to Tab.6, iterative depth prediction improves the performance in both baseline and our network

Backbone	MAE↓	RMSE↓	RMSElog↓	AbsRel↓
ResNet18	0.2309	0.3957	0.0771	0.1195
ResNet34	0.2044	0.3661	0.0687	0.1041
ResNet50	**0.1870**	**0.3410**	0.0664	**0.0984**
ResNet101	0.1911	0.3481	**0.0654**	0.0992

Table 7: Ablation studies about different ResNet backbones.

Method	Parameters	GPU Mem	FPS
BiFuse	253.1M	4003M	0.9
UniFuse	30.26M	1221M	31
SliceNet	75.3M	1911M	13
HoHoNet	49.5M	1487M	52
Ours (Baseline)	52.5M	2136M	19
Ours (Simple)	86.4M	2376M	12
Ours	87.0M	2378M	11

Table 8: Model complexity and computation efficiency

as it decomposes different scales of depth regression and improves the process of gradient backpropagation. Meanwhile, our ACDConv efficiently extracts features for more precise depth estimation and works independently with iterative depth prediction.

ResNet Backbone   Finally, we test different ResNet backbones in our ACDNet. As shown in Tab.7, the network performance gradually improves with the increasing of the backbone complexity. However, using the ResNet101 backbone to build the network is time-consuming and produces overfitting. Considering network performance and overhead, we select ResNet50 as our backbone in the experiments.

Model Complexity   We compare the model complexity and computational efficiency with previous methods, and all the results are derived from inferring a 512×1024 image. Compared with *Baseline*, the dilated convolution groups introduced in *Simple* increase $33.9M$ parameters and $240M$ memories and reduce the FPS from 19 to 12 as shown in Tab.8. Against the data in Tab.2, simply increasing model parameters does not play a fundamental role in improving the performance. While the channel-wise fusion modules in *Ours* scarcely influence the model complexity and computational efficiency but improve results substantially.

Conclusion

In this work, we first propose the adaptively combined dilated convolution to replace the regular convolution to well extract the features from panorama images in ERP. Then we construct the ACDNet to estimate depth maps with monocular panorama images, which outperforms the SOTA approaches in quantitative metrics and visual effects.

Our experiments show that the convolutions with extended receptive fields contribute to panoramic depth estimation. Moreover, the experiments with adaptive channelwise fusion strategy also express that obtaining different latent shapes of convolution kernels in different feature channels is better than explicitly deforming convolution kernels

at different latitudes. That is worth further researching for panorama images in the future. In addition, we will study our ACDConv in other existing depth prediction models and its effects on other various panoramic image tasks such as image classification and semantic segmentation.

References

Armeni, I.; Sax, S.; Zamir, A. R.; and Savarese, S. 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *CoRR* abs/1702.01105.

Chang, A. X.; Dai, A.; Funkhouser, T. A.; Halber, M.; Nießner, M.; Savva, M.; Song, S.; Zeng, A.; and Zhang, Y. 2017. Matterport3D: Learning from RGB-D Data in Indoor Environments. In *3DV*, 667–676. IEEE Computer Society.

Chen, H.; Li, K.; Fu, Z.; Liu, M.; Chen, Z.; and Guo, Y. 2021. Distortion-Aware Monocular Depth Estimation for Omnidirectional Images. *IEEE Signal Process. Lett.* 28: 334–338.

Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans. Pattern Anal. Mach. Intell.* 40(4): 834–848.

Cheng, X.; Wang, P.; Zhou, Y.; Guan, C.; and Yang, R. 2020. Omnidirectional Depth Extension Networks. In *ICRA*, 589–595. IEEE.

Cohen, T.; Geiger, M.; Köhler, J.; and Welling, M. 2017. Convolutional Networks for Spherical Signals. *CoRR* abs/1709.04893.

Cohen, T. S.; Geiger, M.; Köhler, J.; and Welling, M. 2018. Spherical CNNs. In *ICLR*. OpenReview.net.

Coors, B.; Condurache, A. P.; and Geiger, A. 2018. SphereNet: Learning Spherical Representations for Detection and Classification in Omnidirectional Images. In *ECCV (9)*, volume 11213 of *Lecture Notes in Computer Science*, 525–541. Springer.

Eder, M.; Moulon, P.; and Guan, L. 2019. Pano Popups: Indoor 3D Reconstruction with a Plane-Aware Network. In *3DV*, 76–84. IEEE.

Eder, M.; Price, T.; Vu, T.; Bapat, A.; and Frahm, J. 2019. Mapped Convolutions. *CoRR* abs/1906.11096.

Fang, Z.; Chen, X.; Chen, Y.; and Gool, L. V. 2020. Towards Good Practice for CNN-Based Monocular Depth Estimation. In *WACV*, 1080–1089. IEEE.

Fernandez-Labrador, C.; Fácil, J. M.; Pérez-Yus, A.; Demonceaux, C.; Civera, J.; and Guerrero, J. J. 2020. Corners for Layout: End-to-End Layout Recovery From 360 Images. *IEEE Robotics Autom. Lett.* 5(2): 1255–1262.

Fu, H.; Gong, M.; Wang, C.; Batmanghelich, K.; and Tao, D. 2018. Deep Ordinal Regression Network for Monocular Depth Estimation. In *CVPR*, 2002–2011. IEEE Computer Society.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE Computer Society.

Jiang, H.; Sheng, Z.; Zhu, S.; Dong, Z.; and Huang, R. 2021. UniFuse: Unidirectional Fusion for 360° Panorama Depth Estimation. *IEEE Robotics Autom. Lett.* 6(2): 1519–1526.

Jin, L.; Xu, Y.; Zheng, J.; Zhang, J.; Tang, R.; Xu, S.; Yu, J.; and Gao, S. 2020. Geometric Structure Based and Regularized Depth Estimation From 360 Indoor Imagery. In *CVPR*, 886–895. IEEE.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In *ICLR (Poster)*.

Laina, I.; Rupprecht, C.; Belagiannis, V.; Tombari, F.; and Navab, N. 2016. Deeper Depth Prediction with Fully Convolutional Residual Networks. In *3DV*, 239–248. IEEE Computer Society.

Lee, M.; Hwang, S.; Park, C.; and Lee, S. 2021. EdgeConv with Attention Module for Monocular Depth Estimation. *CoRR* abs/2106.08615.

Lee, S.; Lee, J.; Kim, D.; and Kim, J. 2020. Deep Architecture With Cross Guidance Between Single Image and Sparse LiDAR Data for Depth Completion. *IEEE Access* 8: 79801–79810.

Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S. E.; Fu, C.; and Berg, A. C. 2016. SSD: Single Shot MultiBox Detector. In *ECCV (1)*, volume 9905 of *Lecture Notes in Computer Science*, 21–37. Springer.

Ma, H.; Ding, Y.; Wang, L.; Zhang, M.; and Li, D. 2018. Depth Estimation from Monocular Images Using Dilated Convolution and Uncertainty Learning. In *PCM (2)*, volume 11165 of *Lecture Notes in Computer Science*, 13–23. Springer.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 8024–8035.

Pintore, G.; Agus, M.; Almansa, E.; Schneider, J.; and Gobbetti, E. 2021. SliceNet: Deep Dense Depth Estimation From a Single Indoor Panorama Using a Slice-Based Representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11536–11545.

Su, Y.; and Grauman, K. 2017. Learning Spherical Convolution for Fast Features from 360° Imagery. In *NIPS*, 529–539.

Sun, C.; Sun, M.; and Chen, H.-T. 2021. HoHoNet: 360 Indoor Holistic Understanding With Latent Horizontal Features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2573–2582.

Tateno, K.; Navab, N.; and Tombari, F. 2018. Distortion-Aware Convolutional Filters for Dense Prediction in Panoramic Images. In *ECCV (16)*, volume 11220 of *Lecture Notes in Computer Science*, 732–750. Springer.

Tian, Y.; Zhang, Q.; Ren, Z.; Wu, F.; Hao, P.; and Hu, J. 2019. Multi-Scale Dilated Convolution Network Based Depth Estimation in Intelligent Transportation Systems. *IEEE Access* 7: 185179–185188.

Wang, F.; Yeh, Y.; Sun, M.; Chiu, W.; and Tsai, Y. 2020a. BiFuse: Monocular 360 Depth Estimation via Bi-Projection Fusion. In *CVPR*, 459–468. IEEE.

Wang, N.; Solarte, B.; Tsai, Y.; Chiu, W.; and Sun, M. 2020b. 360SD-Net: 360° Stereo Depth Estimation with Learnable Cost Volume. In *ICRA*, 582–588. IEEE.

Wang, T.; Huang, H.; Lin, J.; Hu, C.; Zeng, K.; and Sun, M. 2018. Omnidirectional CNN for Visual Place Recognition and Navigation. In *ICRA*, 2341–2348. IEEE.

Yang, Y.; Wang, Y.; Zhu, C.; Zhu, M.; Sun, H.; and Yan, T. 2021. Mixed-Scale Unet Based on Dense Atrous Pyramid for Monocular Depth Estimation. *IEEE Access* 9: 114070–114084.

Yu, F.; Koltun, V.; and Funkhouser, T. A. 2017. Dilated Residual Networks. In *CVPR*, 636–644. IEEE Computer Society.

Zeng, W.; Karaoglu, S.; and Gevers, T. 2020. Joint 3D Layout and Depth Prediction from a Single Indoor Panorama Image. In *ECCV (16)*, volume 12361 of *Lecture Notes in Computer Science*, 666–682. Springer.

Zhang, J.; Yue, H.; Wu, X.; Chen, W.; and Wen, C. 2020. Densely Connecting Depth Maps for Monocular Depth Estimation. In *ECCV Workshops (4)*, volume 12538 of *Lecture Notes in Computer Science*, 149–165. Springer.

Zheng, J.; Zhang, J.; Li, J.; Tang, R.; Gao, S.; and Zhou, Z. 2020. Structured3D: A Large Photo-Realistic Dataset for Structured 3D Modeling. In *ECCV (9)*, volume 12354 of *Lecture Notes in Computer Science*, 519–535. Springer.

Zhou, K.; Wang, K.; and Yang, K. 2020. PADENet: An Efficient and Robust Panoramic Monocular Depth Estimation Network for Outdoor Scenes. In *ITSC*, 1–6. IEEE.

Zioulis, N.; Karakottas, A.; Zarpalas, D.; Alvarez, F.; and Daras, P. 2019. Spherical View Synthesis for Self-Supervised 360° Depth Estimation. In *3DV*, 690–699. IEEE.

Zioulis, N.; Karakottas, A.; Zarpalas, D.; and Daras, P. 2018. OmniDepth: Dense Depth Estimation for Indoors Spherical Panoramas. In *ECCV (6)*, volume 11210 of *Lecture Notes in Computer Science*, 453–471. Springer.