Distortion-Aware Convolutional Filters for Dense Prediction in Panoramic Images

Keisuke Tateno[1,2], Nassir Navab[1,3], and Federico Tombari[1]

[1] CAMP - TU Munich, Germany
[2] Canon Inc. , Japan
[3] Johns Hopkins University , USA

Abstract. There is a high demand of 3D data for 360° panoramic images and videos, pushed by the growing availability on the market of specialized hardware for both capturing (e.g., omni-directional cameras) as well as visualizing in 3D (e.g., head mounted displays) panoramic images and videos. At the same time, 3D sensors able to capture 3D panoramic data are expensive and/or hardly available. To fill this gap, we propose a learning approach for panoramic depth map estimation from a single image. Thanks to a specifically developed distortion-aware deformable convolution filter, our method can be trained by means of conventional perspective images, then used to regress depth for panoramic images, thus bypassing the effort needed to create annotated panoramic training dataset. We also demonstrate our approach for emerging tasks such as panoramic monocular SLAM, panoramic semantic segmentation and panoramic style transfer.

1 Introduction

The availability of 360° panoramic visual data is quickly increasing thanks to the availability on the market of a new generation of cheap and compact omni-directional cameras: to name a few, Ricoh Theta, Gear360, Insta360 One. At the same time, there is also a growing demand of utilizing such visual content within 3D panoramic displays as provided by head mounted displays (HMDs) and new smartphone apps, dictated by emerging applications in the field of virtual reality (VR) and gaming. Nevertheless, the great majority of currently available panoramic content is just monoscopic, since available hardware has no means to associate depth or geometry information to the acquired RGB data. This naturally limits the sense of 3D when experiencing such content, even if the current hardware could already exploit 3D content, since almost all HMDs feature a stereoscopic display.

Therefore, the ability to acquire 3D data for panoramic images is strongly desired from both a hardware and an application standpoint. Nevertheless, acquiring depth from a panoramic video or image is not an easy task. Conversely to the case of conventional perspective imaging, where there are off-the-shelf, cheap and lightweight 3D sensors (*e.g.* Intel RealSense, Orbbec Astra), consumer 3D omni-directional cameras have not yet been developed. Current devices for obtaining 360° panoramic RGB-D images rely on a set of depth cameras (*e.g.* the Matterport camera[4]), a laser scanner (*e.g.* FARO[5]),

[4] https://matterport.com
[5] https://www.faro.com

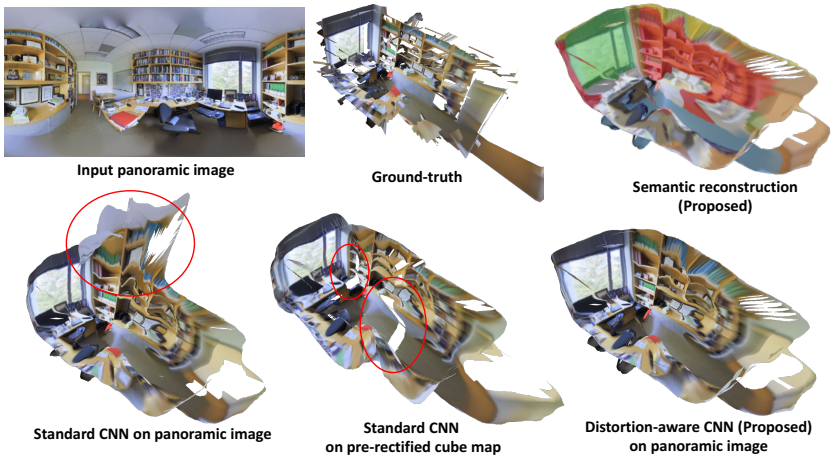

Fig. 1. From a single input equirectangular image (top left), our method exploits distortion-aware convolutions to notably reduce the distortions in depth prediction that affect conventional CNNs (bottom row). Top right: the same idea can be used to predict semantic labels, so to obtain panoramic 3D semantic segmentation from a single image.

or a mobile robotic setup (*e.g.* the NavVis trolley[6]). All these solutions are particularly expensive, require long set-up times and are not suited to mobile devices. Additionally, most of these solutions require static working conditions and cannot deal with dynamic environments, since the devices incrementally scan the surroundings either via mechanical rotation or being pushed around.

Recently a research trend has emerged aiming at depth prediction from a single RGB image. In particular, the use of convolutional neural networks (CNNs) [15, 4, 5] in an end-to-end fashion has proved the ability to regress dense depth maps at a relatively high resolution and with good generalization accuracy, even in the absence of monocular cues to drive the depth estimation task. With our work, we aim to explore the possibility of predicting depth information from monoscopic 360° panoramic image using a learned approach, which would allow obtaining depth information based simply on low-cost omni-directional cameras. One main challenge to accomplish this goal is represented by the need of extensive annotations for training depth prediction, which would still require the aforementioned high-cost, impractical solutions based on 3D panoramic sensors. Instead, if we could exploit conventional perspective images for training a panoramic depth predictor, this would be greatly beneficial for reducing the cost of annotations and for training under a variety of conditions (outdoor/indoor, static/dynamic, *etc.*), by exploiting the wealth of publicly available perspective datasets.

With this motivation, our goal is to develop a learning approach which trains on perspective RGB images and regresses 360° panoramic depth images. The main problem is represented by the distortions caused by the equirectangular representation: indeed, when projecting the spherical pixels to a flat plane, the image gets remarkably distorted

[6] http://www.navvis.com/

especially along the y axis. This distortion leads to significant error in depth prediction, as shown in Fig. 1 (bottom row, left). A simple but partial solution to this problem is represented by rectification. Since 360° panoramic images cannot be rectified to a single perspective image due to the limitations of the field of view of the camera model, they are usually rectified using a collection of 6 perspective images, each associated to a different direction, *i.e*. a representation known as *cube map projection* [8]. However, such representation includes discontinuities at each image border, despite the panoramic image being continuous on those regions. As a consequence, the predicted depth also shows unwanted discontinuities, as shown in Fig. 1 (bottom row, middle), since the receptive field of the network is terminated on the cube map's borders. For this problem, Su *et al*. [29] proposed a method for domain adaptation of CNNs from perspective image to equirectangular panoramic image. Nevertheless, their approach relies on feature extraction specifically aimed at object detection, hence it does not easily extend to dense prediction tasks such as depth prediction and semantic segmentation.

We propose to modify the network's convolutions by leveraging geometrical priors for the image distortion, by means of a novel *distortion-aware* convolution that adapts its receptive field by deforming the shape of the convolutional filter according to the distortion and projection model. Thus, these modified filters can compensate for the image distortions directly during the convolutional operation, so to rectify the receptive field. This allows employing different distortion models for training and testing a network: in particular, the advantage is that panoramic depth prediction can be trained by means of standard perspective images. An example is shown in Fig. 1 (bottom row, right), highlighting a notable reduction of the distortions with respect to standard convolutions. We demonstrate the domain adaptation capability for the depth prediction task between rectified perspective images and equirectangular panoramic images on a public panoramic image benchmarks, by replacing the convolutional layers of a state-of-the-art architecture [15] with the proposed distortion-aware convolutions. Moreover, we also test our approach for semantic segmentation and obtain 360° semantic 3D reconstruction from a single panoramic image (see Fig. 1, top right). Finally, we show examples of application of our approach for tasks such as panoramic monocular SLAM and panoramic style transfer.

2 Related works

Depth prediction from single image There is an increasing interest towards depth prediction from single image thanks to the recent advances in deep learning. Classic depth prediction approaches employ hand-crafted features and probabilistic graphical models [11][17] to yield regularized depth maps, usually by over constraining the scene geometry. Recently developed deep convolutional architectures significantly outperformed previous methods in terms of depth estimation accuracy [15][4][5][25][24][18][16]. Compared with such supervised method, unsupervised depth prediction based on stereo images was also proposed [7][14]. This is particularly suitable for scenarios where accurate dense range data is difficult to obtain, *e.g*. outdoor and street scenes.

Deformation of the Convolutional Unit Approaches to deform the shape of the convolutional operator to improve the receptive field of a CNN have been recently ex-

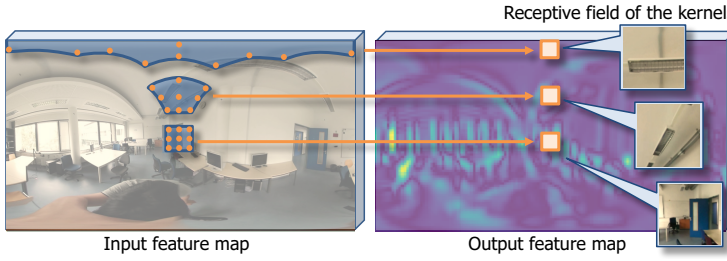

Fig. 2. The key concept behind the distortion-aware convolution is that the sampling grid is deformed according to the image distortion model, so that the receptive field is rectified.

plored [13][12][3]. Jeon *et al*. propose a convolution unit with learned offsets to obtain better receptive field for object classification, by learning fixed offsets for feature sampling on each convolution. Dai *et al*. propose a more dynamically deformable convolution unit where the image offsets are learned through a set of parameters [3]. Henriques *et al*. propose a warped convolution to make the network invariant to general spatial transformations such as translation and scale changes or 2D and 3D rotation [10]. Su *et al*. propose a method to learn specific convolution kernel along each horizontal scanline so to adapt a CNN trained on perspective images to the equirectangular domain [29]. Each convolutional kernel is retrained so that the error between the output of the kernel in the perspective image and that in the equirectangular image is minimized. Although they aim to solve a similar problem as our work, their domain adaptation approach focuses specifically on object detection and classification, so it cannot be directly applied to dense prediction tasks such as depth prediction and semantic segmentation. Additionally, their method needs to re-train each network individually to adapt to the equirectangular image domain, even though the image distortion coefficients would remain exactly the same.

3D shape recovery from single 360 image Approaches to recover 3D shape and semantic from a single equirectangular image by geometrical fusion have been explored in [27][26]. Yang *et al*. propose a method to recover the 3D shape from a single equirectangular image by analyzing vertical and horizontal line segments and superpixel facets in the scene by imposing geometric constraints [27]. Xu *et al*. propose a method to estimate the 3D shape of indoor spaces by combining surface orientation estimation and object detection [26]. Both algorithms don't use machine learning, and rely on the Manhattan world assumption, hence these methods can deal only with indoor scenes that present vertical and horizontal lines. Therefore these methods cannot be applied to scenes that present an unorganized structures, such as outdoor environments.

3 Distortion-aware CNN for depth prediction

In this section, we formulate the proposed distortion-aware convolution operator. We first introduce the basic operator in Sec. 3.1. Then in Sec. 3.2 we describe how to compute an adaptive spatial sampler within the distortion-aware convolution according to

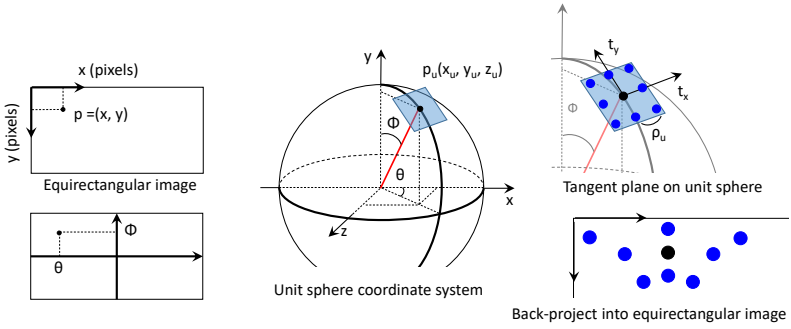

Fig. 3. Overview of computation of the adaptive sampling grid for equirectangular image. Each pixel p in the equirectangular image is transformed into unit sphere coordinates, then the sampling grid is computed on the tangent plane in unit sphere coordinates, finally the sampling grid is back-projected into equirectangular image to determine the location of the distorted sampling grid.

the equirectangular projection. Subsequently, in Sec. 3.3 we illustrate the architecture of our dense prediction network with distortion-aware convolutions for depth prediction and semantic segmentation.

3.1 Distortion-aware Convolution

In the description of our convolution operator, for the sake of clarity, we consider only the part regarding the 2D spatial convolution out of the 4D convolutional tensor, and drop the notation and description regarding the additional dimensions related to the number of channels and batch size. The 2D convolution operation is carried out following two steps: first, features are sampled by applying a regular grid $\mathcal{R}$ on the input feature map f_l at layer l, then the sum of a neighborhood of features weighted by w is computed. The sampling grid $\mathcal{R}$ defines the receptive field size and scale. In case of a standard 3×3 filter, the grid is simply defined as

$$\mathcal{R} = \{(-1,-1),(-1,0),...,(1,0),(1,1)\} \, . \tag{1}$$

A generic 2D spatial location on a feature map, grid or image is denoted as $p = (x\,(p)\,, y\,(p))$, *i.e.* x and y are the operators returning, respectively, the horizontal and vertical coordinate of the location p.

For each location p on the input feature map f_l, each output feature map element f_{l+1} is computed as

$$f_{l+1}(p) = \sum_{r \in \mathcal{R}} w(r) \cdot f_l(p + r) \tag{2}$$

where r enumerates the pixel relative location in $\mathcal{R}$.

In the distortion-aware convolution, the sampling grid $\mathcal{R}$ is transformed by means of a function $\delta(p, r)$ which computes a distorted neighborhood of pixel locations according to the image distortion model. In this case, (2) becomes

$$f_{l+1}(p) = \sum_{r \in \mathcal{R}} w(r) \cdot f_l\left(p + \delta\left(p, r\right)\right) \, . \tag{3}$$

By adaptively deforming the sampling grid according to the distortion function $\delta(p, r)$, the receptive field gets rectified, as shown in Fig. 2. Details regarding how to compute $\delta(p, r)$ according to the distortion model are given in Sec. 3.2.

The pixel location computed by means of $\delta(p, r)$ is mostly fractional, thus (3) is computed via bilinear interpolation as

$$f_{l+1}(p) = \sum_{q \in \aleph(\tilde{p})} G(q, \tilde{p}) f_l(q) \tag{4}$$

where $\tilde{p}$ is the fractional pixel location obtained by means of the distortion function $\delta(p, r)$, *i.e.* $\tilde{p} = p + \delta(p, r)$, and $\aleph(\tilde{p})$ denotes the four integer spatial locations adjacent to $\tilde{p}$. Moreover, $G(\cdot, \cdot)$ represents the bilinear interpolation kernel, *i.e.*

$$\begin{aligned} &G(q, p) \\ &= \max(0, 1 - |x(q) - x(p)|) \max(0, 1 - |y(q) - y(p)|). \end{aligned} \tag{5}$$

Importantly, in case of undistorted perspective images, the result of the convolution as defined in (3) is the same as that of the regular convolution in (2).

3.2 Sampling grid transformation via unit sphere coordinate system.

Here, we describe how to compute the distorted pixel location $\delta(p, r)$ from the pixel location p and the relative location of the sampling grid $r = (x(r), y(r)) \in \mathcal{R}$. Fig. 3 illustrates the whole set of transformations applied across different coordinate systems.

First, the image coordinates of a point p on the equirectangular image (x, y) are transformed to a longitude and a latitude in the spherical coordinate system $p_s = (\theta, \phi)$ as

$$\theta = (x - \frac{w}{2}) \frac{2\pi}{w} \tag{6}$$

$$\phi = (\frac{h}{2} - y) \frac{\pi}{h} \tag{7}$$

where w and h are, respectively, the width and height of the input image in pixels.

Then, the latitude and longitude (θ, ϕ) are converted to the unit sphere coordinate system $p_u = (x_u, y_u, z_u)$ according to the following relations:

$$p_u = \begin{bmatrix} x_u \\ y_u \\ z_u \end{bmatrix} = \begin{bmatrix} \cos(\phi) \sin(\theta) \\ \sin(\phi) \\ \cos(\phi) \cos(\theta) \end{bmatrix} \tag{8}$$

Subsequently, the tangent plane in the unit sphere coordinate system around the pixel location of p_u, *i.e.* $t_u = (t_x, t_y)$, is computed. To this aim, the horizontal and vertical direction vectors t_x, t_y of the tangential plane can be obtained by means of the upper vector of the unit sphere coordinate system $\upsilon = (0, 1, 0)$ as

$$t_x = |\upsilon \times p_u| \tag{9}$$

$$t_y = |p_u \times t_x| \tag{10}$$

where $\times$ represents the cross product of two vectors.

At this point, we note that the projection of the image on such tangent plane represents the rectified image around the pixel location on the original equirectangular image p. Hence, the desired set of distorted pixel locations on the original image $\hat{p}$ can be obtained via back-projection of the neighboring locations on the tangent plane t_u sampled via a regular grid to the equirectangular image coordinate system. This sampling grid, denoted as r_{sphere}, is computed using the two axes of the tangent plane t_x, t_y and the relative element locations on the original sampling grid $r = (x(r), y(r)) \in \mathcal{R}$. Hence, each element of the grid can be defined as

$$r_{sphere} = \rho_u \cdot (t_x \cdot r\,(x) + t_y \cdot r\,(y)) \tag{11}$$

where ρ_u represents the spatial resolution (*i.e.*, distance between elements) on the unit sphere coordinate system corresponding to the resolution of the initial equirectangular image. The resolution equivalent to 1 pixel on the equirectangular image can be computed as:

$$\rho_u = \tan\left(\frac{2\pi}{w}\right) . \tag{12}$$

Although not discuss here but interesting in perspective, while this resolution is equivalent to no dilation of the sampling kernel, a generic dilation of the kernel can be obtained by increasing the value of ρ_u, this leads to the definition of atrous convolutions [28] for panoramic images.

Each location on the tangent plane related to the sampling grid element r_{sphere} is then computed as

$$p_{u,r} = p_u + r_{sphere} . \tag{13}$$

Finally, each element $p_{u,r} = (x_{u,r}, y_{u,r}, z_{u,r})$ is back-projected to the equirectangular image domain by using the inverse function of the aforementioned coordinate transformations, first by going through the spherical coordinate system, *i.e.* inverting (8)

$$\theta_r = \begin{cases} \tan^{-1}(\frac{z_{u,r}}{x_{u,r}}) & (\text{if } x_{u,r} >= 0) \\ \tan^{-1}(\frac{z_{u,r}}{x_{u,r}}) + \pi & (\text{otherwise}) \end{cases} \tag{14}$$

$$\phi_r = sin^{-1}(y_{u,r}) \tag{15}$$

then by landing on the original 2D equirectangular image domain

$$x(r) = (\frac{\theta_r}{2\pi} + \frac{1}{2})w \tag{16}$$

$$y(r) = (\frac{1}{2} - \frac{\phi_r}{\pi})h . \tag{17}$$

The previously defined function $\delta(p, r)$ computes the relative coordinates $x(r)-x(p), y(r)-y(p)$. Since these offsets are constant given the image distortion model, they can be computed once and stored for later use. In the case of equirectangular images (and differently from fish-eye images), since the distortions are constant over the same horizontal location, only a set of $h * |R|$ offsets needs to be stored ($|R|$ being the number of elements in the grid/filter). Also important to note, from a geometrical point of view, the distortion-aware convolution as defined above is equivalent to the convolutional operation applied on the tangent plane in the unit sphere coordinate system.

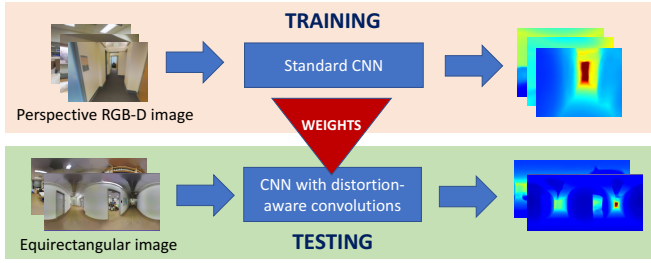

Fig. 4. A major advantage of the proposed approach is that standard convolutional architectures can be used with common datasets for perspective images to train the weights. At test time, the weights are transferred on the same architecture with distortion-aware convolutional filters so to process equirectangular images. Although the figure report the case of depth prediction, we apply the same strategy for the semantic segmentation task.

3.3 CNN architecture for dense prediction task

In general, the distortion-aware convolution operator can be applied to any type of CNN architecture by replacing the standard convolutional operator. In this work, we build our architecture by modifying the fully convolutional residual network (FCRN) model proposed in [15], given the competitive results obtained for both depth prediction and semantic segmentation. The downsampling part of the FCRN architecture is based on ResNet-50 [9], and initialized with pre-trained weights from ImageNet [20], while the upsampling part replaces the fully connected layers originally in ResNet-50 with a set of up-sampling residual blocks composed of unpooling and convolutional layers. The loss function is based on the reverse Huber function [15], while weights are optimized via back-propagation and Stochastic Gradient Descent (SGD).

As for the modifications that need to be applied on the network, each spatial convolution unit in FCRN is replaced with a distortion-aware convolution. The pixel shuffler units such as the fast up-convolution unit that was proposed in [15] to increase computational efficiency are replaced with a normal unpooling and convolution, since pixel shuffling in fast-up convolution assumes that pixel neighbors are always consistent, while feature sampling in distortion-aware convolution does not keep pixel neighbor consistency. Additionally, for the unpooling layers, we replace max unpooling with average unpooling, *i.e.* taking the average value of the two nearest neighbors to fill the empty entries. Indeed, max unpooling, which uses zeros to fill the empty entries, cannot be used with the fractional sparse sampling used by distortion-aware convolution, since interpolation with zeros inevitably leads to artifacts in the output feature map. Additionally, to obtain pixel-wise semantic segmentation labels rather than depth values, the final layer is modified so to have as many output channels as the number of classes, while the loss is the cross-entropy function.

This paradigm allows us to train the network by leveraging commonly used datasets with annotations for perspective images, and to test using as input equirectangular panoramic images. Indeed, the weights are exactly the same between the standard version of the network and its distortion-aware counterpart. This idea is depicted in Fig. 4. This is a major advantage in the case of panoramic images due to the aforementioned

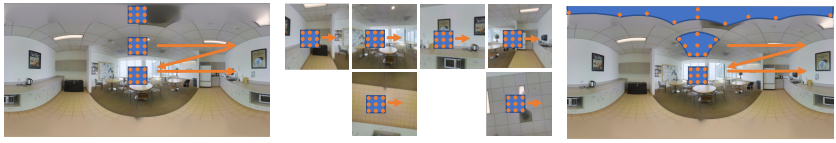

Fig. 5. Compared methods in our experimental evaluation: (a) Standard convolution on equirectangular image, (b) Standard convolution on 6 rectified images via cube map projection, (c) Proposed distortion-aware convolution on equirectangular image.

Fig. 6. Example of equirectangular image with/without inpainting and extracted rectifed perspective images.

limitations of public datasets with dense annotations for depth prediction and semantic segmentation tasks.

4 Evaluation

This section provides an experimental evaluation of our method for the tasks of depth prediction (Sec. 4.2) and semantic segmentation (Sec. 4.3) on equirectangular 360° panoramic images. We compare it both quantitatively and qualitatively to standard convolution on equirectangular images as well as cube-map rectification, *i.e.* the standard method to rectify 360° spherical images, as shown in Fig. 5. In addition, we show the application of panoramic depth prediction to outdoor data and to panoramic monocular SLAM. Finally we show the generalization of our distortion-aware convolution to a different task (i.e., panoramic style transfer) and a different CNN architecture (i.e., VGG). The supplementary material include further qualitative evaluation.

4.1 Experimental setup

For the implementation of our distortion-aware convolution and dense prediction network we use TensorFlow[7]. We train on a single NVIDIA Geforce GTX 1080 with 8GB of GPU memory. The weights of the encoding layers of the FCRN architecture are pre-trained on the NYU Depth v2 dataset [21] while the modified layers of the up-convolutions (average unpooling and convolutions) are initialized as random filters sampled from a normal distribution with zero mean and 0.01 variance. As described in Sec. 3.3, the network is trained on rectified perspective RGB images to predict the corresponding depth maps using standard convolutions, then it is tested on equirectangular images by means of distortion-aware convolutions. As benchmark for testing, we use

[7] https://www.tensorflow.org

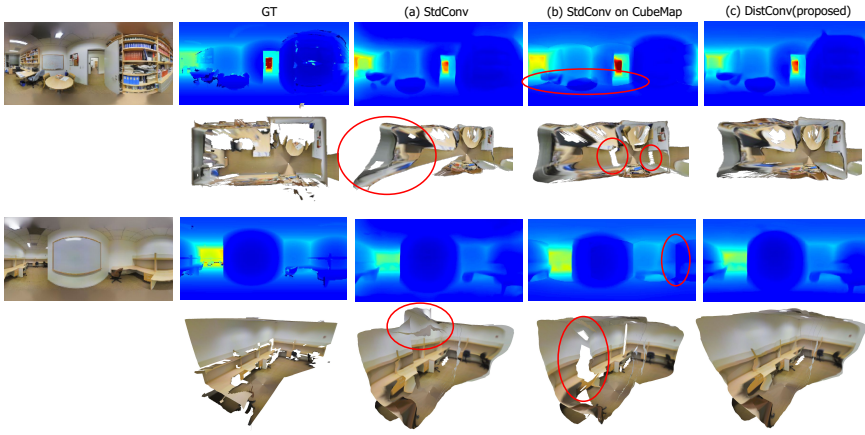

Fig. 7. Qualitative comparison of depth prediction on Stanford 2D-3D-S dataset [1]. Red circles highlight artifacts due to distortions induced by the standard convolutional model (a) and by the CubeMap representation (b) that are instead solved by our approach (c).

the *Stanford 2D-3D-S* dataset [1], that provides equirectangular 360° panoramic images with depth and semantic labels as ground-truth annotations. The dataset consists of 1412 images, captured with the Matterport sensor[8], where the official split includes 1040 images for training and 372 for testing.

Since the images on this dataset lack color nearby polar regions, they are filled in with zeros (see Fig. 6(a)). To avoid biasing the network during training, we apply an inpainting algorithm [23] as shown in Fig. 6(b). To create perspective images for training, first we extract images with limited field of view along different directions from the original 360° panoramic image. Directions are sampled on a 20° interval along the vertical axis (yaw rotation) and on a 15° interval along the horizontal axis (pitch rotation). Then, we rectify them into a standard perspective view as shown in Fig. 6(c). These rectified perspective images are created by mapping pixels from the equirectangular projection to the perspective projection [8]. The total number of training image is $216320 = 1040 \times 16 \times 13$. Note that the depth image of a 360° panoramic image stores the distance with respect to the direction from the camera center position to the point, and not along the z-axis of the camera coordinate system (front view direction) as it usually occurs with standard perspective depth maps. This is due to the fact that if a camera has a field of view larger than 180°, it could not define negative depths along the front view direction (it would be 0 or less). Hence, the depth map of the extracted and rectified perspective images is also encoded using the distance values instead of the depth values. We train the FCRN model with standard convolutions with a batch size of 16 for approximately 20 epochs. The starting learning rate is 0.01 for all layers, which we gradually reduce every 6-8 epochs, when we observe plateaus; momentum is 0.9. The rectified perspective images in training are rescaled to 308×228 pixels, while the equirectangular images used for testing are rescaled to 960×480 pixels, so that spatial resolution of 1° of the view angle is comparable between training and testing.

[8] https://matterport.com

Table 1. (1) Comparison in terms of depth prediction accuracy on Stanford 2D-3D benchmark dataset, and (2) Comparison on Stanford 2D-3D benchmark dataset, trained by NYU depth dataset v2.

	(1) Stanford 2D-3D-S			(2) NYU depth dataset		
	rel [m]	rms [m]	log10	rel [m]	rms [m]	log10
(a) StdConv	0.201	0.395	0.094	0.604	0.631	0.188
(b) StdConv on CubeMap	0.220	0.371	**0.0818**	0.669	0.692	0.195
(c) DistConv (ours)	**0.176**	**0.369**	0.0829	**0.517**	**0.578**	**0.171**

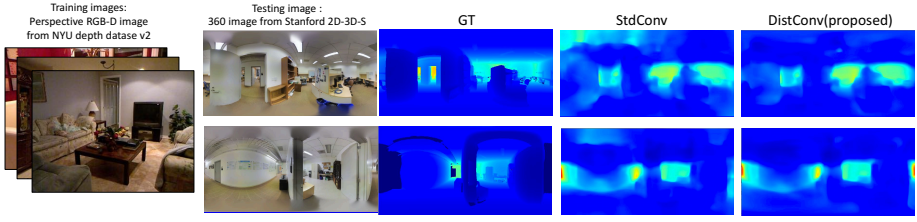

Fig. 8. Examples of depth prediction on Stanford 2D-3D-S dataset predicted by the network trained by NYU depth dataset v2.

4.2 Panoramic depth prediction

Table 1 reports the accuracy of depth prediction computed using standard error metrics as proposed in previous works [15][5][4], *i.e*. the relative error (rel), the root mean square error (rms) and the log10 scaled error (log10) between the ground-truth depth and the predicted one. Given the results in the table, we can conclude that our method outperforms related methods. Notably, in terms of relative error, which is particularly sensitive to small errors, our method shows a remarkably improved performance compared to the others. In the metrics on log10 and rms, our method and cubemap rectification show comparable result. However, in terms of relative error, cubemap is worse than the other two. The cause of this can be determined by looking at the qualitative results shown in Fig. 7, that reports both the predicted depth maps as well as the top-view reconstruction of the three evaluated method, and compares them to the ground-truth. The result of standard convolution is quite inaccurate (visible in particular in the top-view image), due to the discontinuity along image borders and the distortions along polar regions. The result of cubemap does not show such shape deformations, but there are depth ”jumps” near the image borders on each cube map, as visible from the predicted depth map and the top-view image. This is due to the limited field of view of each image of the cube map, which limits the receptive field of the CNN on such regions.

To complement previous results, we also demonstrate how our distortion-aware convolution can be tested on equirectangular images while trained on benchmark perspective datasets. This experiment also shows the generalization capabilities of our approach to adapt to different datasets between training and testing. In this case, the network is trained on the benchmark NYU depth dataset V2 [21] and tested on Stanford 2D-3D-S [1]. We train the FCRN model in a similar manner as described in 4.2, but using only

Table 2. Comparison in terms of semantic segmentation accuracy of each category on Stanford 2D-3D benchmark dataset. The accuracy is computed as Intersect over Union (%).

	ceiling	floor	wall	column	beam	window	door	table	chair	bookcase	sofa	board	clutter	total
(a) StdConv	60.82	78.01	54.85	0	0	40.11	13.08	34.55	32.45	44.91	0	46.65	18.84	32.63
(b) CubeMap	61.32	72.72	**61.77**	0	0.21	36.97	**15.45**	**37.54**	33.48	**48.50**	0	48.34	**23.42**	33.82
(c) DistConv	**61.56**	**83.40**	57.17	0	**0.376**	**42.65**	13.85	37.38	**35.41**	47.17	0	**50.85**	19.52	**34.56**

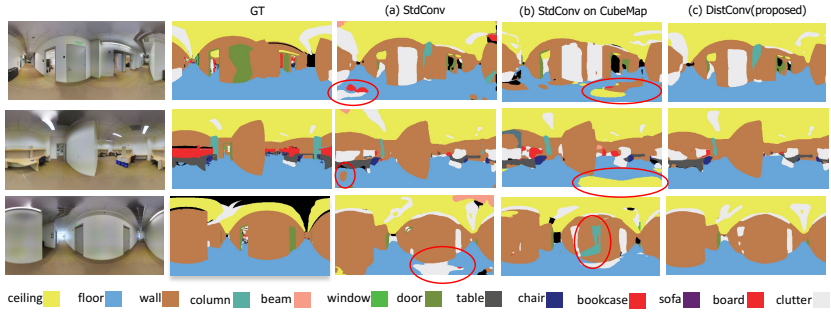

Fig. 9. Qualitative comparison of semantic segmentation on Stanford 2D-3D-S dataset [1]. Red circles highlight errors on polar regions and borders of the CubeMap model that are not present in our distortion-aware approach.

data from the NYU depth dataset. During training the perspective images are rescaled to 160×128 pixels, while the equirectangular images used for testing are rescaled to 960×480 pixels, so that the spatial resolution on 1° of the view angle is comparable between training and testing. An example of a training image is shown on the left of Fig. 8. The quantitative results are shown in (2) of table 1. Our method outperforms standard convolution, although the prediction accuracy is decreased due to the different domain of the scene. Also the qualitative results are shown in Fig. 8. Generally, the result by standard convolution tends to fail on polar regions in the predicted depth map. On the other hand, our proposed method can predict correctly on such regions.

4.3 Panoramic semantic segmentation

We evaluate our distortion-aware convolution for the task of panoramic semantic segmentation. The semantic labels in Stanford 2D-3D-S dataset consist of 13 semantic classes. We carry out an evaluation by comparing the same 3 methods as done for the depth prediction experiment. Table 2 reports the accuracy of semantic segmentation, computed as the mean of class-wise intersection over union (mIoU), *i.e.* using the same metrics used in related work for semantic segmentation [19][30]. As shown in the table, our method shows better accuracy compared to the standard convolution and the cube map approach. In particular, our method significantly improves the accuracy of "floor" class because such a structure can be often found around polar regions on the equirectangular image, i.e. where strong distortions are usually present, which are typically problematic for standard convolution. The overall accuracy for other classes such

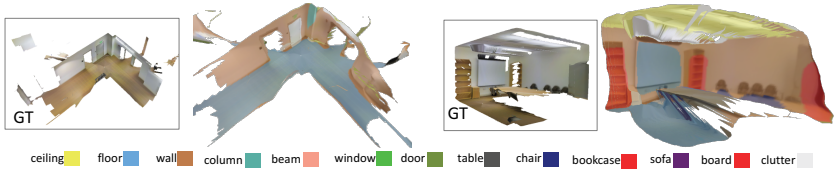

Fig. 10. Qualitative result of our depth prediction and semantic segmentation from monoscopic 360° panoramic image.

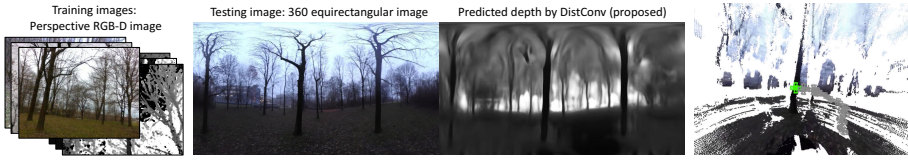

Fig. 11. Left and middle: Predicted depth in outdoor scene trained by perspective images obtained from an Xtion pro depth camera. Right: an example of reconstructed scene and estimated poses by CNN-SLAM-360

as ”window” and ”chair” is also better. From the qualitative results in Fig. 9, we can see that standard convolution yields segmentation errors especially near polar regions. Also, incorrect segmentation artifacts can be seen from the outcome of cube map, caused by the field-of-view limitation on each cube map image. On the other hand, our method reports higher accuracy within these regions. We also show the combined result of depth prediction and semantic segmentation in Fig. 10,left. The semantic reconstruction result is inferred by means of a single monoscopic 360° image. Remarkably, our method allows to jointly reconstruct and semantically segment the entire scene around the camera from a single image, which would not be possible neither by standard depth prediction nor by SLAM or structure-from-motion.

4.4 Outdoor scenes and panoramic monocular SLAM

To complement previous results, we show the performance of our method in outdoor settings. Since our method does not rely on any geometric assumption, such as the Manhattan world assumption used by [27][26], it can be applied also on outdoor scenes. In this case, we use a pre-trained network on NYU v2 and Stanford 2D-3D-S datasets, then fine-tuned by means of 1200 RGB-D images obtained by Xtion pro live (shown on the left of Fig.11). The network is tested on the equirectangular image acquired via Insta360 One omni-directional camera[9]: the predicted map is shown in Fig.11, middle. Notably, our method can predict depth from outdoor scenes by pre-training on benchmark datasets and fine-tuning by means of a consumer depth camera.

We also demonstrate the extension to panoramic monocular SLAM based on monoscopic 360° panoramic sequences. To this goal, we have borrowed the idea of CNN-SLAM [22], that refines CNN-based depth prediction with depth estimates from monocular SLAM, yielding camera pose estimation and fused 3D reconstruction. To apply

[9] https://www.insta360.com/product/insta360-one/

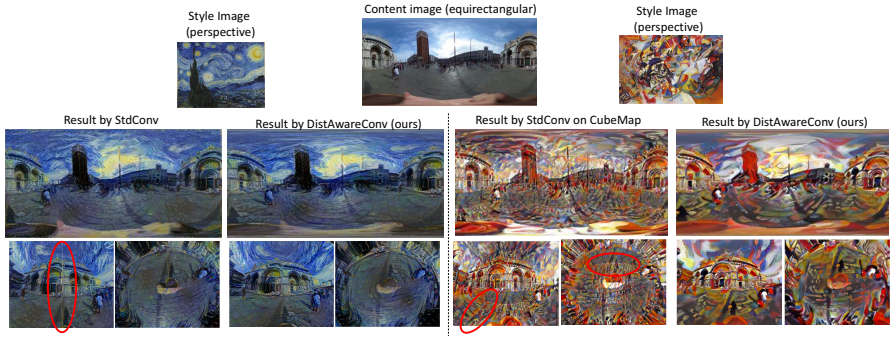

Fig. 12. Application of our distortion-aware convolution for panoramic style transfer.

CNN-SLAM on equirectangular data, we introduce multiple pin-hole camera models similar to the omni-directional approach in [2]. An example of reconstruction and estimated camera poses is shown on Fig. 11, right. Additional qualitative results are included in the supplementary material.

4.5 Application to panoramic style transfer

Being our distortion-aware convolution general purpose in terms of tasks and independent from the specific network architecture, we apply our proposed convolution to a different task named panoramic style transfer, i.e. an extension to equirectangular panoramic images of the style transfer on perspective images proposed in [6]. Here we do not employ the FCRN network but the modified VGG architecture proposed in [6], where the part of the network used to encode the input image content is modified by replacing standard convolutions with distortion-aware ones. Since the style images that we use are normal perspective images, the network layers which encode the style image rely on the original convolutions. The middle row in Fig.12 shows the result of style transfer while the bottom row shows the perspective image projected from the style transfered equirectangular image. As the red highlights show, some border and discontinuity can be seen on the results by standard convolution and the result on Cube map, because the style transfer by standard convolution does not consider the distortion and continuity of equirectangular image. On the other hand, the projected images from our method do not show such discontinuities and appear more natural.

5 Conclusion

The proposed distortion-aware convolution proved to be effective compared to standard convolution as well as the CubeMap representation on two dense prediction tasks such as depth prediction and semantic segmentation. We also showed the successful application to different architectures (FCRN and VGG), purely perspective training sets (NYU v2) and further tasks such as panoramic style transfer. Future work includes extending our approach to different distortion models such as equidistance projection and equisolid angle projection for fisheye lens and different prediction tasks such as object detection or instance segmentation in equirectangular images.

References

1. Armeni, I., Sax, A., Zamir, A.R., Savarese, S.: Joint 2D-3D-Semantic Data for Indoor Scene Understanding. ArXiv e-prints (2017)
2. Caruso, D., Engel, J., Cremers, D.: Large-scale direct slam for omnidirectional cameras. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2015)
3. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: International Conference on Computer Vision (ICCV), 2017 (2017)
4. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: In Proc. Int. Conf. Computer Vision (ICCV) (2015)
5. Eigen, D., Puhrsch, C., Fergus, R.: Prediction from a single image using a multi-scale deep network. In: Proc. Conf. Neural Information Processing Systems (NIPS) (2014)
6. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR) (2016)
7. Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: In Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR) (2017)
8. Greene, N.: Environment mapping and other applications of world projections. IEEE Computer Graphics and Applications (1986)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. Proc. Conf. Computer Vision and Pattern Recognition (CVPR) (2016)
10. Henriques, J.F., Vedaldi, A.: Warped convolutions: Efficient invariance to spatial transformations. In: International Conference on Machine Learning (ICML) (2017)
11. Hoiem, D., Efros, A., Hebert, M.: Geometric context from a single image. In: In Computer Vision and Pattern Recognition (CVPR) (2005)
12. Jaderberg, M., Simonyan, K., Zisserman, A., kavukcuoglu, k.: Spatial transformer networks. In: Advances in Neural Information Processing Systems (NIPS) (2015)
13. Jeon, Y., Kim, J.: Active convolution: Learning the shape of convolution for image classification. In: Proc. Conf. Computer Vision and Pattern Recognition (CVPR) (2017)
14. Kuznietsov, Y., Stuckler, J., Leibe, B.: Semi-supervised deep learning for monocular depth map prediction. In: In Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR) (2017)
15. Laina, I., Rupprecht, C., Belagiannis, V., Tombari, F., Navab, N.: Deeper depth prediction with fully convolutional residual networks. In: IEEE International Conference on 3D Vision (3DV) (arXiv:1606.00373) (October 2016)
16. Li, B., Shen, C., Dai, Y., den Hengel, A.V., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In: Proc. Conf. Computer Vision and Pattern Recognition (CVPR). pp. 1119–1127 (2015)
17. Liu, B., Gould, S., Koller, D.: Single image depth estimation from predicted semantic labels. In: In Computer Vision and Pattern Recognition (CVPR) (2010)
18. Liu, F., Shen, C., Lin, G.: Deep convolutional neural fields for depth estimation from a single image. In: Proc. Conf. Computer Vision and Pattern Recognition (CVPR). pp. 5162–5170 (2015)
19. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: In Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR) (2015)
20. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015)

21. Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)
22. Tateno, K., Tombari, F., Laina, I., Navab, N.: Cnn-slam: Real-time dense monocular slam with learned depth prediction. In: In Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR) (2017)
23. Telea, A.: An image inpainting technique based on the fast marching method. Journal of graphics tools 9.1 (2004)
24. Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L.: Towards unified depth and semantic prediction from a single image. In: Proc. Conf. Computer Vision and Pattern Recognition (CVPR). pp. 2800–2809 (2015)
25. Xu, D., Ricci, E.: Multi-scale continuous crfs as sequential deep networks for monocular depth estimation. In: In Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR) (2017)
26. Xu, J., Stenger, B., Kerola, T., Tung, T.: Pano2cad: Room layout from a single panorama image. In: IEEE Winter Conference on Applications of Computer Vision (WACV) (2017)
27. Yang, H., Zhang, H.: Efficient 3d room shape recovery from a single panorama. In: Proc. Int. Conf. on Computer Vision and Pattern Recognition (CVPR) (2016)
28. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: ICLR (2016)
29. Yu-Chuan, S., Kristen, G.: Flat2sphere: Learning spherical convolution for fast features from 360 imagery. In: Proc. Conf. on Neural Information Processing Systems (NIPS) (2017)
30. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: In Proc. Int. Conf. Computer Vision and Pattern Recognition (CVPR) (2017)