Signal, Image and Video Processing (2022) 16:643–650
https://doi.org/10.1007/s11760-021-02003-3

ORIGINAL PAPER

Semantic segmentation of outdoor panoramic images

Semih Orhan[1] · Yalin Bastanlar[1]

Received: 4 June 2021 / Revised: 4 June 2021 / Accepted: 1 August 2021 / Published online: 14 August 2021

Abstract
Omnidirectional cameras are capable of providing 360° field-of-view in a single shot. This comprehensive view makes them preferable for many computer vision applications. An omnidirectional view is generally represented as a panoramic image with equirectangular projection, which suffers from distortions. Thus, standard camera approaches should be mathematically modified to be used effectively with panoramic images. In this work, we built a semantic segmentation CNN model that handles distortions in panoramic images using equirectangular convolutions. The proposed model, we call it UNet-equiconv, outperforms an equivalent CNN model with standard convolutions. To the best of our knowledge, ours is the first work on the semantic segmentation of real outdoor panoramic images. Experiment results reveal that using a distortion-aware CNN with equirectangular convolution increases the semantic segmentation performance (4% increase in $mIoU$). We also released a pixel-level annotated outdoor panoramic image dataset which can be used for various computer vision applications such as autonomous driving and visual localization. Source code of the project and the dataset were made available at the project page (https://github.com/semihorhan/semseg-outdoor-pano).

Keywords Semantic segmentation · Panoramic images · Omnidirectional vision · Convolutional neural networks

1 Introduction

Semantic segmentation is a fundamental and challenging problem of computer vision. It is the task of assigning semantic labels to each pixel in images. Many computer vision applications benefit from it, such as pedestrian detection [6,16], autonomous vehicles [22,26], pose estimation [19,27] and remote sensing [13,24]. In the last decade, convolutional neural networks (CNNs) have become the best among the approaches of semantic segmentation with their capability of learning fine and coarse details.

Unlike conventional cameras with narrow field-of-view (FOV), omnidirectional cameras are capable of capturing 360° view with a single shot. Due to their wide FOV, they have gained popularity in many computer vision application areas from autonomous vehicles to augmented reality. This led to an increasing amount of effort to adapt various computer vision tasks, especially object detection and semantic segmentation, for 360° imagery.

A full omnidirectional view (360° horizontally and 180° vertically) is generally represented as a panoramic image with equirectangular projection. Coordinates are proportional to latitude and longitude of the sphere, i.e., unit distance in horizontal or vertical direction in the image corresponds to a fixed amount of angular coverage (Fig. 1). It heavily suffers from distortions toward the top and bottom which causes objects to appear differently. This is a challenge for computer vision methods which were optimized only with standard FOV images.

To tackle distortions in panoramic images, several methods have been proposed [4,10,23,25], where the distortion is modeled by an explicit implementation of convolution kernel. By doing so, the convolution is performed, not on regular grid coordinates, but on coordinates shifted by offsets calculated regarding spherical distortion.

In this work, we introduce a version of UNet [20], a semantic segmentation CNN, where we replaced standard convolution layers with equirectangular convolutions [10] so that it can alleviate the effects of distortion in panoramic images. Although there are previous works [12,28] on semantic segmentation of panoramic images, these studies are

✉ Semih Orhan
semihorhan@iyte.edu.tr

Yalin Bastanlar
yalinbastanlar@iyte.edu.tr

[1] Department of Computer Engineering, Izmir Institute of Technology, Izmir, Turkey

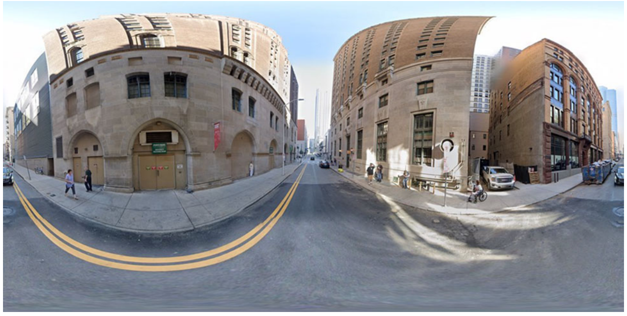

Fig. 1 A panoramic image with equirectangular projection. Source: Google Street View

focused on either indoor environments [12] or synthetic panoramic images [28]. In our work, we investigate the semantic segmentation performance on real panoramic outdoor images for the first time. We report CNN performance on panoramic outdoor images when the distortions are corrected with equirectangular convolutions. We also evaluate various transfer learning options.

In addition, we release a pixel-level semantically annotated panoramic outdoor dataset, called as CVRG-Pano, especially for visual-based localization and autonomous driving tasks. Images belong to the urban and suburban areas of Pittsburgh and downloaded from Google Street View. Dataset consists of 600 images. We labeled 20 separate semantic classes which are grouped into 7 categories. Images were fine-level pixel annotated with an effort of approximately 500 man-hours. CVRG-Pano shares most of the classes with Cityscapes [5] so that it can be used for training on 360° images after initialized with Cityscapes weights.

We summarize the contributions of our work below:

- We propose a semantic segmentation CNN that handles panoramic image distortions explicitly by using equirectangular convolutions (UNet-equiconv). We test its performance on an outdoor panoramic real image dataset for the first time.
- We release a pixel-level annotated outdoor panoramic image dataset for semantic segmentation. It is the first of its kind. It can be used for a wide range of computer vision applications, and we believe it will be beneficial to the computer vision community.

The remainder of this paper is structured as follows. We review the related work in Sect. 2. We describe the proposed model and explain equirectangular convolution in Sect. 3. We explained our dataset in Sect. 4. Experimental results are given in Sect. 5 which is followed by conclusions in Sect. 6.

2 Related work

In the past years, deep neural networks have demonstrated outstanding performance on semantic segmentation. The literature review given below covers deep learning-based methods. We categorize the previous studies into two: the first group worked with perspective (standard FOV) images, whereas the second group focused on panoramic images with or without handling distortions.

2.1 Semantic segmentation of perspective images

One of the first CNN-based models, designed for semantic segmentation, is fully convolution networks (FCN) [15]. In [15], Long et al. discarded the classification layer of well-known CNN models, such as AlexNet and VGG, and converted fully connected layers to convolution layers. By doing so, they could work with variable sizes of inputs. Encoded features are upsampled by applying skip connection and bilinear interpolation on the different depths of the networks. Noh et al. [17] introduced DeconvNet, which is a pioneer encoder–decoder model. The first part of the model consists of convolution and pooling layers that encode the features. The second part of the network upsamples the feature maps by using deconvolution (transpose convolution) and unpooling layers. Another encoder–decoder network is SegNet [1], which is similar to DeconvNet [17]. The encoder part consists of the same number of convolutional layers as VGG has, and the decoder part consists of deconvolution and unpooling layers. The key novelty of SegNet is max-pooling indexes are used in the decoder part, which helped to reduce total number of parameters of the model. Ronneberger et al. proposed UNet [20]. In UNet, entire features are upsampled using bilinear interpolation and concatenated at the decoder part of the network using skip connections. There are recent better-performing semantic segmentation CNN models, such as DeepLabv3+ [3]; however, in this paper we preferred to build a UNet-based model due to its effectiveness and ease of implementation.

2.2 Semantic segmentation of panoramic images

There is a recent body of research focused on how to handle distortions while applying CNNs on panoramic images. The main idea is that we should move rectangular convolution kernels on the sphere representations rather than the panoramic images. Coors *et al.* [4] introduced spherical convolution. They modified standard convolutional layer by calculating offsets of the grids regarding spherical distortion and presented studies on object detection. Tateno et al. [25] present results on dense depth estimation and semantic segmentation, where CNN was trained with normal FOV images but used to get results from panoramic images with distortion-aware kernels.

Su and Grauman [23] proposed SPHCONV. It learns spherical convolution with training on perspective images. The proposed model does not require an annotated panoramic dataset. However, due to the narrow FOV of perspective cameras, the close objects cannot be fully captured in the training set which leads not performing well for close objects. Another limitation was that parameter size linearly increases with the height of the equirectangular input image.

Deng et al. [8] introduced OOP-net. Apart from the conventional CNN models, OOP-net has overlapping pyramid pooling (OPP) module which explores local and global contextual information at the same time. That study used fisheye cameras, where FOV is 180°, without proposing a method for handling distortions. In [28], Xu *et al.* released a synthetic panoramic outdoor dataset for semantic segmentation. Panoramic synthetic images are obtained from SYNTHIA sequence dataset [21].

Fernandez-Labrador et al. [10] proposed equirectangular convolution for 3D layout estimation. This is a special form of deformable convolution [7], where offsets of the kernel elements are fixed and calculated according to the equirectangular projection. In a follow-up study, Guerrero-Viu et al. [12] introduced an equirectangular version of BlitzNet [9] which is designed for semantic segmentation and object detection. They worked on panoramic indoor dataset and reported that performance increases when the network is trained with panoramic images and equirectangular convolution.

In our work, we follow the distortion handling approach of [10] and introduce an equirectangular version of UNet model, called as UNet-equiconv. Different from previous works, we directly work with an outdoor dataset of real panoramic images. We investigate, for the first time, the effects of transfer learning and distortion handling performance on outdoor panoramic images.

3 Method

3.1 Network architecture

We introduce an equirectangular version of UNet [20], UNet-equiconv, where each convolution layer is replaced with equirectangular convolution (Sect. 3.2) to compensate distortions. The architecture of UNet-equiconv is shown in Fig. 2. Each convolution is followed by batch normalization and rectified linear units (ReLU), which are omitted in the figure for the sake of simplicity. Also, repetitive convolution layers are represented with ‘x.’ For instance ‘x4’ means the same convolution is repeated four times. Output depth is 8, corresponding to 7 semantic categories and one ‘unlabeled’ category.

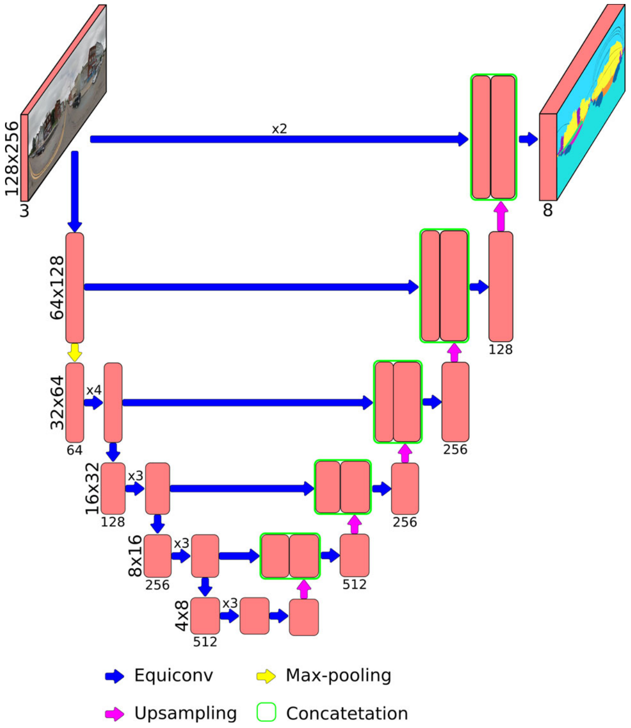

Fig. 2 Architecture of UNet-equiconv

3.2 Equirectangular convolution

Previous CNN-based object detection and semantic segmentation studies showed that when the convolutions are modified to compensate the distortions, accuracy increases ([4,12,23,25]). The main idea is that we should move the convolution kernels on the sphere rather than the panoramic images (Fig. 3). The kernel is rotated and applied along the sphere, and its position is defined by the spherical coordinates (θ and ϕ) of its center. In practice, we use square kernels. Here, we describe how to compute distorted pixel location from the kernel entities for location p on the unit sphere. Figure 3 illustrates the whole set of transformations applied across different coordinate systems.

We follow the steps described in [10] and first define ($u_{0,0}$, $v_{0,0}$) as the corresponding pixel location on the equirectangular image where we apply the convolution operation (i.e., the image coordinate where the center of the kernel is located). Then, these coordinates are transformed to a longitude and a latitude in the spherical coordinate system (Fig. 3b).

$$\theta_{0,0} = \left(u_{0,0} - \frac{W}{2}\right)\frac{360}{W}; \quad \phi_{0,0} = -\left(v_{0,0} - \frac{H}{2}\right)\frac{180}{H} \tag{1}$$

where θ and ϕ are in degrees and W and H are, respectively, the width and height of the equirectangular image in pixels.

Subsequently, the 3D coordinates for every element in the kernel (the tangent plane) is computed (Fig. 3c). When we consider a 3x3 kernel on the equator, kernel element 3D coordinates are:

$$\hat{p}_{ij} = \begin{bmatrix} \hat{x}_{ij} \\ \hat{y}_{ij} \\ \hat{z}_{ij} \end{bmatrix} \tag{2}$$

where i and j are the horizontal and vertical indexes of a kernel element. 3D coordinates change as follows:

$$\hat{p}_{0,0} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \quad \hat{p}_{\pm 1,0} = \begin{bmatrix} \pm \tan \Delta_\theta \\ 0 \\ 1 \end{bmatrix}, \quad \hat{p}_{0,\pm 1} = \begin{bmatrix} 0 \\ \pm \tan \Delta_\phi \\ 1 \end{bmatrix} \tag{3}$$

where Δ_θ and Δ_ϕ are $360/W$ and $180/H$ in degrees, respectively. These correspond to the angles covered by one pixel in the equator of the sphere. When the filter size is larger, angular coverage of kernel also decreases. Although we do not employ, lower resolution kernels can also be defined for wide angles. Readers can find detailed formulation on various kernel resolutions in [10].

We keep the kernel shape on the tangent plane fixed. When applying the filter at a different location (θ,ϕ), we rotate the points to the corresponding point of the sphere. We also project each point onto the sphere surface by normalizing the vectors:

$$p_{ij} = \begin{bmatrix} x_{ij} \\ y_{ij} \\ z_{ij} \end{bmatrix} = R_y(\phi_{0,0}) R_x(\theta_{0,0}) \frac{\hat{p}_{ij}}{|\hat{p}_{ij}|} \tag{4}$$

where $R_a(\beta)$ stands for a rotation matrix of an angle β around a axis.

Finally, the rest of elements are back-projected to the equirectangular image domain (Fig. 3d). First, 3D kernel coordinates are transferred to latitude and longitude angles, which is called as the inverse gnomonic projection:

$$\theta_{ij} = \arctan\left(\frac{x_{ij}}{z_{ij}}\right); \quad \phi_{ij} = \arcsin\left(y_{ij}\right) \tag{5}$$

Then, converted to the original 2D equirectangular image domain:

$$u_{ij} = \left(\frac{\theta_{ij}}{360} + \frac{1}{2}\right) W; \quad v_{ij} = \left(-\frac{\phi_{ij}}{180} + \frac{1}{2}\right) H \tag{6}$$

Equirectangular convolution is a special form of deformable convolution [7], where the convolution is not performed by a grid-like kernel, but learned offsets are added to the kernel locations. In panoramic images, deformation does not have a free form and the offsets in kernel follow a pattern. Thus, the offsets are not learned but computed using the geometry of equirectangular projection. The offsets are constant as the kernel moves horizontally, but they increase as the convolution kernel moves to the poles of the sphere (Fig. 4).

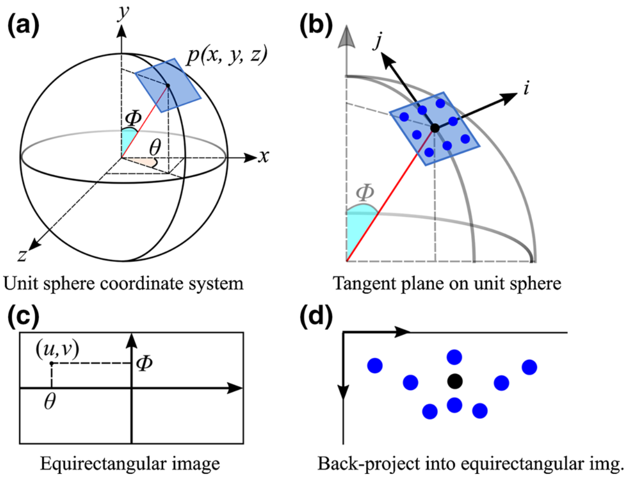
(a)
y
p(x, y, z)
Φ
θ
x
z
Unit sphere coordinate system
(b)
j
i
Φ
Tangent plane on unit sphere
(c)
(u,v)
Φ
θ
Equirectangular image
(d)
Back-project into equirectangular img.

Fig. 3 Distortion-aware convolution. 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, and finally, the sampling grid is backprojected into equirectangular image to determine the location of the distorted sampling grid

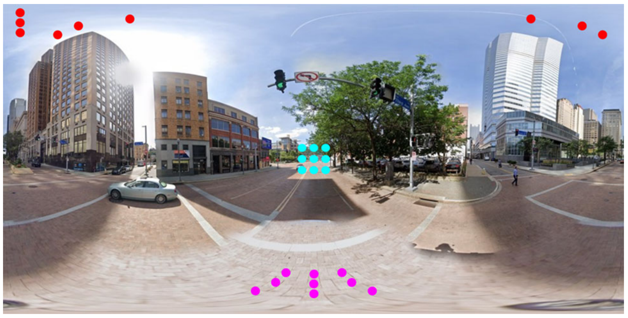

Fig. 4 The offsets of spherical convolution. Three different kernel positions are shown to highlight the differences between offsets. On the equator, kernel maps the neighboring pixels. As we approach to the poles, the deformation gets bigger. When the borders are exceeded, the points are taken from the other side of the 360° image

4 Dataset

Commonly used datasets of semantic segmentation were prepared with standard FOV cameras [2,5,11,14]. Providing 360° view presents important benefits to a wide range of applications. Therefore, we decided to release a pixel-level annotated 360° outdoor panoramic dataset. We hope that our effort will be beneficial to research in various sub-fields such as visual localization and autonomous driving.

Table 1 Details of categorical grouping

Flat	Ground, road, sidewalk, parking
Construction	Building, wall, fence, bridge
Object	Pole, traffic light, traffic sign
Nature	Vegetation, terrain
Sky	Sky
Person	Person
Vehicle	Car, truck, bus, motorcycle, bicycle

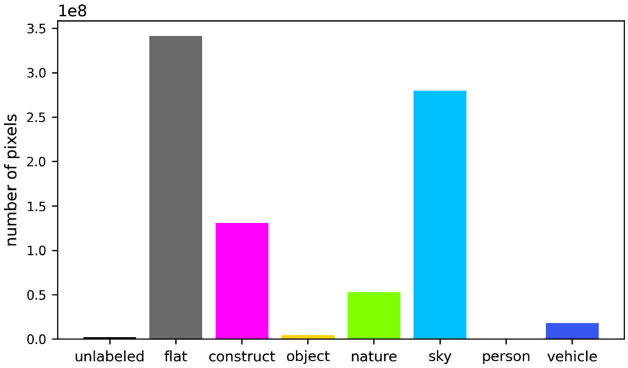

Fig. 5 Number of annotated pixels (y-axis) of each category, and their categorical labels (x-axis)

4.1 Pixel-level annotated dataset

We prepared an outdoor panoramic image dataset with semantic labels, called as CVRG-Pano. Images belong to the urban and suburban parts of Pittsburgh City. They were obtained from Google Street View and downloaded via Street View Download 360 application[1]. The dataset consists of pixel-level annotated 600 images. We manually labeled 20 semantic classes and then grouped them into 7 categories (Table 1) in accordance with the categorization defined in [5]. Figure 5 shows the distribution of categorical dataset. The dataset is divided into three: 446 images for training, 48 images for validation and 76 images for testing. The dataset is available at the project repository[2]. An example image from panoramic dataset and its semantic labels are shown in Fig. 6.

4.2 Automatically generated semantic segmentation dataset

Manual annotation is a labor-intensive task and takes great amount of time. As an alternative, we follow a method to fastly produce semantic masks of panoramic images. First, we generate cubemaps of panoramic images. Then, the semantic mask of each cubemap is generated by a state-of-the-art segmentation CNN model (we used HRNet-OCR [29] trained on Cityscapes). As a final step, we project semantic mask of all cubemaps to panorama. The whole process is illustrated in Fig. 7. This automatically generated panoramic outdoor dataset consists of 504 images. We grouped semantic classes into seven categories conforming to the pixel-level manually annotated dataset (Table 1).

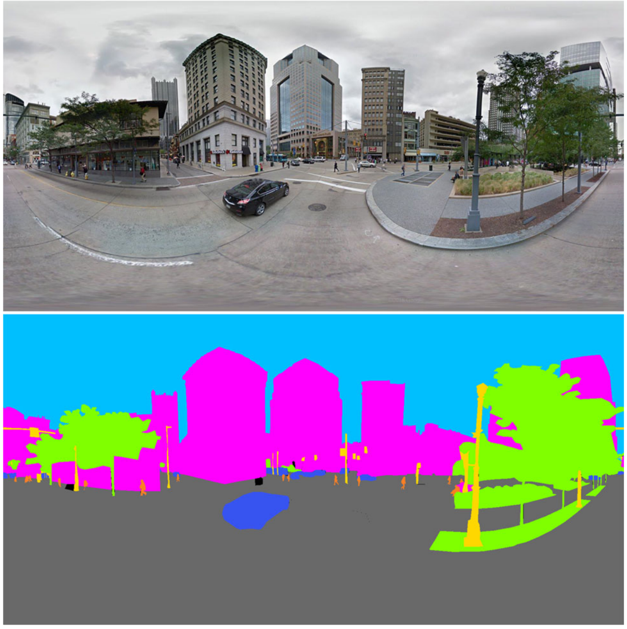

Fig. 6 An example image from CVRG-Pano and its semantic labels

5 Experiments

We conducted several experiments on both datasets with standard and equirectangular convolutions. We mainly report the effects of equirectangular convolution and various weight initializations on the segmentation performance. Moreover, we investigate the usefulness of automatically generated masks. PyTorch [18], version 1.7.1, is used as a deep learning framework, and all models are trained with Nvidia GeForce GTX 1080 GPU.

5.1 Evaluation metric

We evaluated the performance of models using mean intersection over union ($mIoU$) as it is a common evaluation metric (e.g., [5], [12]). In Eq. 7, M represents the total number of classes, A_i is total number of ground truth pixels of

[1] https://iStreetView.com.

[2] https://github.com/semihorhan/semseg-outdoor-pano.

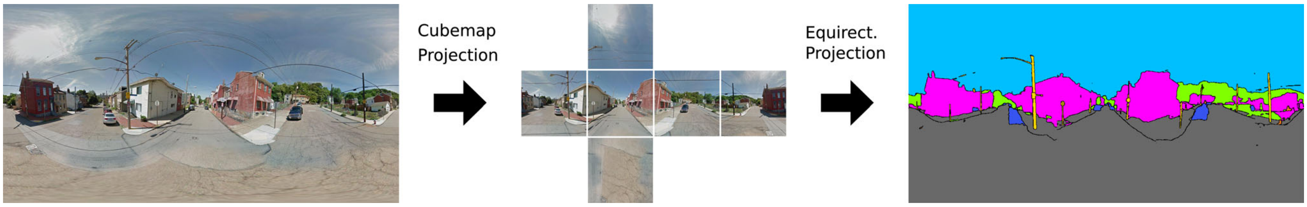

Fig. 7 The processes of automatic semantic mask generation. As a first step, we generate cubemaps of panorama and then generate a mask of each cubemaps by the state-of-the-art CNN model. As the last step, we project a mask of cubemaps to panorama

Table 2 Effect of weight initialization on training with pixel-level annotated panoramic image dataset

	Test *mIoU*
From scratch	0.610
ImageNet	0.634
Cityscapes	0.649

class i, and $\hat{A}_i$ is predicted number of pixels for class i.

$$mIoU = \frac{1}{M} \sum_{i=1}^{M} \frac{A_i \cap \hat{A}_i}{A_i \cup \hat{A}_i} \tag{7}$$

5.2 Weight initialization

Due to having a limited number of training images, we conducted several experiments to maximize the potential of the models. In Table 2, we see the effect of weight initialization on the performance of UNet with standard convolution (UNet-stdconv). Training from scratch and two different transfer learning alternatives are compared. All were trained with pixel-level annotated training set and tested with 76-image test set (cf. Sect. 4.1). We obtained the best result by using pre-trained Cityscapes weights. We can also infer that pre-training on a large dataset (even it is not composed of city images) helps to perform better since the results with the model pre-trained on ImageNet are better than the model trained from scratch.

5.3 Standard versus equirectangular convolution

Here, we investigate how the performance of UNet-stdconv changes when equirectangular convolution is applied (UNet-equiconv). Firstly, all models were trained on Cityscapes [5] dataset, and then they were fine-tuned on the panoramic dataset. Table 3 shows that UNet-equiconv model that uses equirectangular convolution, performs better than UNet-stdconv. Test results support the hypothesis that eliminating distortion at the feature level increases the performance of a CNN model and improves its generalization ability not only on indoor images [12], but also on outdoor images. Qualitative comparison of both models is shown in Fig. 8.

We also observe from Table 3 that the classes that cover larger portions of the image (e.g., sky, flat, construction) have higher $mIoU$ and rare classes have lower $mIoU$. An important contribution of equirectangular convolution is that it helps rare classes (e.g., object and person) significantly. Since these objects cover small areas, especially if they are close to the top or bottom of the image, compensating the distortion may affect the result severely. (First row of Fig. 8 is an example.)

5.4 Usefulness of automatically generated dataset

We also conducted an experiment to investigate the usefulness of automatically generated dataset (cf. Sect. 4.2) for training. First, we trained UNet-stdconv model on Cityscapes [5]. Then, we fine-tuned UNet-stdconv model on automatically generated dataset. Both baseline and fine-tuned models were tested on pixel-level annotated 76 images. According to Table 4, the fine-tuned model performs better (0.626 *mIoU*) than the model which was only trained with perspective images (0.573 *mIoU*). This result shows that the automatically generated dataset, even though there exist wrong labeled pixels due to the automatic mask generation process, helps to adapt weights of the perspective image trained model for panoramic images. Training with a much precise panoramic dataset increases the performance even more (0.649 *mIoU* in Table 2).

6 Conclusions

Panoramic images bring advantages for a wide range of computer vision systems due to having wide FOV, but they suffer from significant distortions. Naive approach would be generating overlapping perspective images and process them with standard methods, but it becomes a computationally expensive solution.

In this work, we propose UNet-equiconv which utilizes equirectangular convolution to alleviate the effect of distortion by explicitly modeling the offsets in the convolution kernel. We conducted several experiments to compare the performances of UNet-stdconv and UNet-equiconv. Results indicate an improvement obtained by using equirectangular

Table 3 Semantic segmentation performance of UNet-stdconv and UNet-equiconv

models	mIoU	flat	construction	object	nature	sky	person	vehicle
UNet-stdconv	0.649	0.960	0.827	0.173	0.787	0.976	0.127	0.693
UNet-equiconv	0.674	0.963	0.841	0.233	0.802	0.978	0.168	0.732

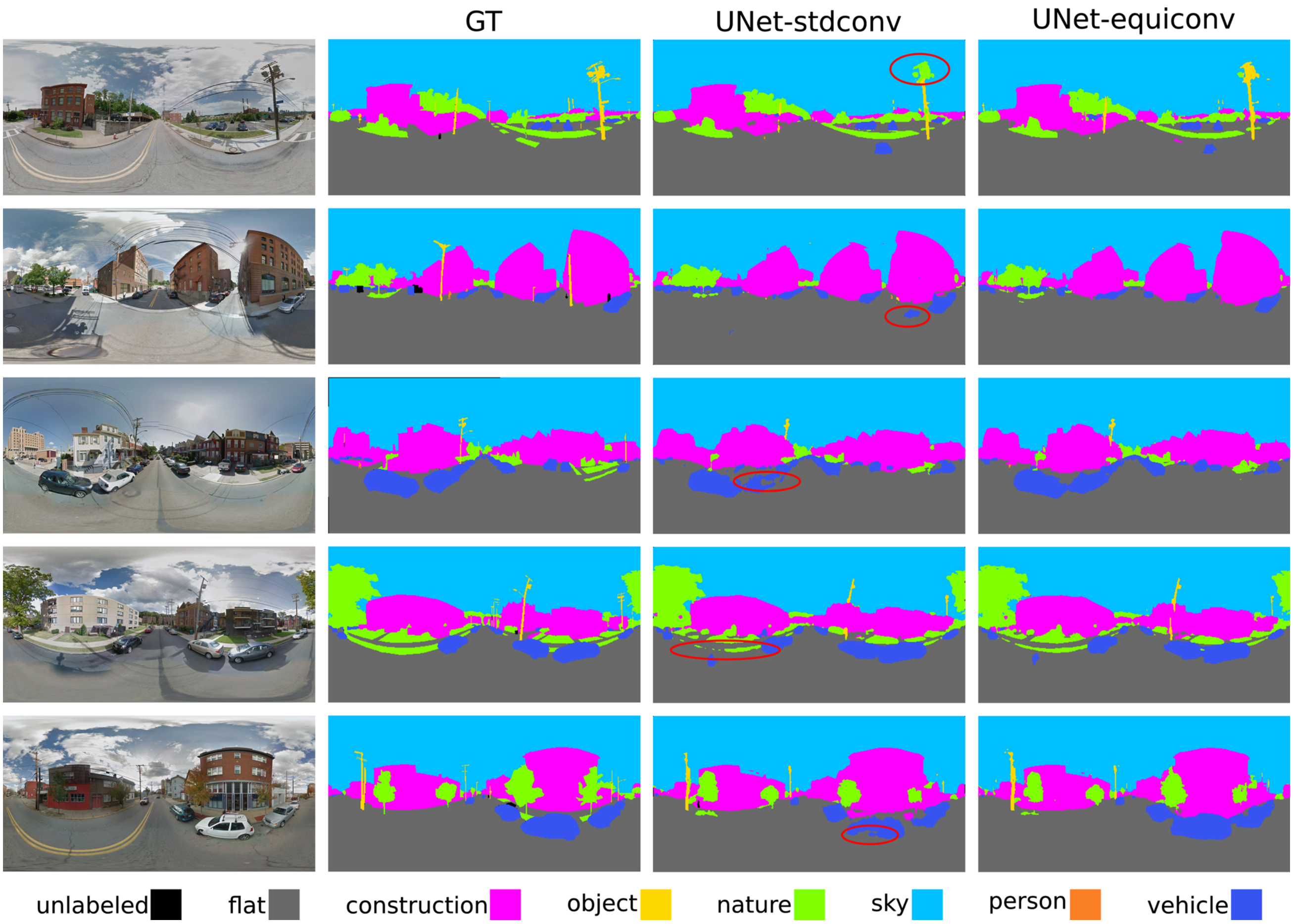

Fig. 8 Qualitative comparison of standard and equirectangular convolution models for semantic segmentation on CVRG-Pano. Red circles highlight some of the errors of standard convolution model that are not present in our distortion-aware approach

Table 4 Effect of fine-tuning UNet-stdconv with automatically generated dataset after pre-trained with Cityscapes

	Test mIoU
Cityscapes	0.573
Auto-pano	0.626

convolution, especially for small size objects which can be represented weakly. Our results are consistent with previous works which utilized equirectangular convolution in indoor environments. We can also say that training with panoramic images increases the performance of a model even if standard convolution is used. We made our semantic segmentation outdoor panoramic dataset publicly available. We hope that the dataset will be useful to the computer vision community.

Acknowledgements This work was supported by the Scientific and Technological Research Council of Turkey (Grant No.120E500)

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **39**(12), 2481–2495 (2017)
2. Brostow, G.J., Fauqueur, J., Cipolla, R.: Semantic object classes in video: a high-definition ground truth database. Pattern Recogn. Lett. **30**(2), 88–97 (2009)
3. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp. 801–818 (2018)
4. Coors, B., Condurache, A.P., Geiger, A.: Spherenet: learning spherical representations for detection and classification in omni-

directional images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 518–533 (2018)

5. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
6. Costea, A.D., Nedevschi, S.: Semantic channels for fast pedestrian detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2360–2368 (2016)
7. Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., Wei, Y.: Deformable convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 764–773 (2017)
8. Deng, L., Yang, M., Qian, Y., Wang, C., Wang, B.: Cnn based semantic segmentation for urban traffic scenes using fisheye camera. In: 2017 IEEE Intelligent Vehicles Symposium (IV), pp. 231–236. IEEE (2017)
9. Dvornik, N., Shmelkov, K., Mairal, J., Schmid, C.: Blitznet: A real-time deep network for scene understanding. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4154–4162 (2017)
10. Fernandez-Labrador, C., Facil, J.M., Perez-Yus, A., Demonceaux, C., Civera, J., Guerrero, J.J.: Corners for layout: end-to-end layout recovery from 360 images. IEEE Robot. Autom. Lett. **5**(2), 1255–1262 (2020)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. IEEE (2012)
12. Guerrero-Viu, J., Fernandez-Labrador, C., Demonceaux, C., Guerrero, J.J.: What's in my room? object recognition on indoor panoramic images. In: 2020 IEEE International Conference on Robotics and Automation (ICRA), pp. 567–573. IEEE (2020)
13. Kampffmeyer, M., Salberg, A.B., Jenssen, R.: Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 1–9 (2016)
14. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European Conference on Computer Vision, pp. 740–755. Springer (2014)
15. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
16. Mao, J., Xiao, T., Jiang, Y., Cao, Z.: What can help pedestrian detection? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3127–3136 (2017)
17. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1520–1528 (2015)
18. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: an imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703 (2019)
19. Peng, S., Liu, Y., Huang, Q., Zhou, X., Bao, H.: Pvnet: Pixel-wise voting network for 6dof pose estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4561–4570 (2019) Network for 6dof Pose Estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4561–4570 (2019)
20. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–241. Springer (2015)
21. Ros, G., Sellart, L., Materzynska, J., Vazquez, D., Lopez, A.M.: The synthia dataset: a large collection of synthetic images for semantic segmentation of urban scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3234–3243 (2016)
22. Siam, M., Gamal, M., Abdel-Razek, M., Yogamani, S., Jagersand, M., Zhang, H.: A comparative study of real-time semantic segmentation for autonomous driving. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 587–597 (2018)
23. Su, Y.C., Grauman, K.: Learning spherical convolution for fast features from 360° imagery. In: NIPS (2017)
24. Sun, W., Wang, R.: Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with dsm. IEEE Geosci. Remote Sens. Lett. **15**(3), 474–478 (2018)
25. Tateno, K., Navab, N., Tombari, F.: Distortion-aware convolutional filters for dense prediction in panoramic images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 707–722 (2018)
26. Teichmann, M., Weber, M., Zöllner, M., Cipolla, R., Urtasun, R.: Multinet: real-time joint semantic reasoning for autonomous driving. In: 2018 IEEE Intelligent Vehicles Symposium (IV), pp. 1013–1020 (2018). https://doi.org/10.1109/IVS.2018.8500504
27. Wong, J.M., Kee, V., Le, T., Wagner, S., Mariottini, G.L., Schneider, A., Hamilton, L., Chipalkatty, R., Hebert, M., Johnson, D.M., et al.: Segicp: integrated deep semantic segmentation and pose estimation. In: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 5784–5789. IEEE (2017)
28. Xu, Y., Wang, K., Yang, K., Sun, D., Fu, J.: Semantic segmentation of panoramic images using a synthetic dataset. In: Artificial Intelligence and Machine Learning in Defense Applications, vol. 11169, p. 111690B. International Society for Optics and Photonics (2019)
29. Yuan, Y., Chen, X., Wang, J.: Object-contextual representations for semantic segmentation. arXiv preprint arXiv:1909.11065 (2019)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.