 DOI: 10.3217/978-3-85125-869-1-03

Efficient Instance Segmentation of Panoramic Images of Indoor Scenes

Werner Bailer and Hannes Fassold

Abstract— **This paper addresses the issue of efficient 2D instance segmentation of 360° images of indoor scenes. In particular, we study the use of equirectangular convolutions and the impact of different approaches to handle wrap-around areas. We consider the use of Mollweide projection as a representation for performing segmentation, and we provide a toolchain to prepare the Matterport panoramic images for use in workflows designed for COCO-style annotated datasets. The results show no significant differences between using regular and equirectangular convolutions. While the Mollweide projection allows for segmentation of otherwise missed objects, the overall results do not outperform analysis on equirectangular projection.**

I. INTRODUCTION

In many application areas (e.g., interior design, furniture retailing or renovation), communication with a customer or future user during the planning and design phase is crucial to select the right products and configurations. Making this communication process effective saves costs, avoids later modifications, and results in providing tailored solutions and higher customer satisfaction. Augmented Reality (AR) has the potential to make these communication processes highly effective and provide a better experience for the customer. However, AR content needs to be created by experts from the respective domains, who often lack IT and media skills, and shall provide a lightweight AR experience for the customer. Current AR authoring solutions are quite complex and require manually creating scenes or rely on objects prepared with even more complex applications (e.g. CAD). In order to facilitate this process, a simple capture process (e.g., using consumer grade 360° cameras) and intelligent scene understanding tools are needed.

One important component is segmenting and classifying the relevant objects such as furniture in interior scenes. In particular, we aim to perform instance segmentation for indoor scenes in single panoramas of rooms. This shall also be possible on consumer hardware with limited processing capabilities. In order to process the 360° images, we aim to avoid training or fine-tuning models specifically for 360° data. This is motivated by the fact that annotated datasets for object segmentation on panoramic images are very scarce. Due to the efficiency requirements, performing the analysis on separate viewports of the 360° image is not feasible.

The authors are with DIGITAL – Institute for Information and Communication Technologies at JOANNEUM RESEARCH, Graz, Austria, {firstname.lastname}@joanneum.at

The research leading to these results has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 951900 ATLANTIS ("Authoring tool for indoor augmented and diminished reality experiences"). The authors thank Nikolaos Zioulis from CERTH ITI for improving the Matterport data preparation script, and Hermann Fürntratt for help with the COCO annotation script.

The contributions of this paper are: (i) we study the use of equirectangular convolutions and the impact of handling the wrap-around areas, (ii) we consider the use of Mollweide projection as a representation for performing the segmentation, and (iii) we provide a toolchain to prepare the Matterport panoramic images for use in workflows designed for COCO-style annotated datasets.

The rest of this paper is organized as follows. Section II discusses related work and Section III presents the approaches that were investigated. Section IV discusses the evaluation (including dataset preparation) and the obtained results, and Section V concludes the paper.

II. RELATED WORK

Impressive progress has been made in instance segmentation of indoor environments represented as point clouds. Such point clouds can be obtained from capturing the scene with multiple views or depth sensors. Recent approaches such as PointGroup [16], 3D-SIS [14] and 3D-MPA [10] show good performance on benchmarks such as ScanNet [7]. However, in many consumer application scenarios, depth information is not available, and thus 2D approaches are required.

We are thus interested in an efficient and reliable 2D instance segmentation approach. A well known approach is Mask R-CNN [12], a two stage instance segmentation based on Faster R-CNN. Masklab [5] is a further evolution of this type of approaches. In terms of efficiency, single stage approaches are preferable. Recent methods showing good performance on benchmark datasets include SOLO v2 [24], Yolact++ [2], proposal free instance segmentation [15] and SipMask [3].

We aim to apply instance segmentation to 360° images. Different approaches to handle this issue have been proposed in literature. One group of methods requires specific training on 360° images or at least fine-tuning. This can be done by adapting early layers of a pretrained network to work on equirectangular images, which is proposed in [22] and tested for object detection using VGG and Faster R-CNN. [6] follow a similar approach with SphereNet, learning a network adjusted to equirectangular inputs. The use of icosahedral Snyder equal-area (ISEA) projections is proposed in [8] and results in significant improvement for semantic indoor segmentation on the SUMO dataset.

To avoid the need for specifically training the network on panoramic data, [25] perform segmentation on multiple stereographic projections. Equirectangular convolutions are proposed in [23] as a convolution kernel for equirectangular images that adjusts the input values to the image positions,

including handling of wrap-around. A similar approach is proposed in [9], with generalised convolutions that use a mapping function. That paper analyses different mapping functions and proposes mapping to a geodesic grid. Equirectangular convolutions have been recently used for indoor semantic segmentation [11], and the authors report small improvements over processing the equirectangular image with standard convolutions.

While a number of approaches for handling 360° images in CNNs have been proposed, many of them require some kind of training or fine-tuning, which limits the practical application. Using specific types of convolutions is reported to improve performance (at least slightly) in some papers, but most of the work deals with object detection rather than segmentation. We are thus interested to study the impact of these choices in our applications, as well as the use of the Mollweide projection, which to the best of our knowledge has not yet been investigated for this purpose.

III. Investigated approaches

Due to the lightweight implementation and the potential to run the method also on a mobile device, we select Yolact++ [2] as the basis of our work. We implement two approaches for processing 360° images with Yolact++: the first one is to integrate equirectangular convolutions, and the second is to transform input images using the Mollweide projection.

For both equirectangular and Mollweide projected images we also optionally extend the image to wrap-around the seam of the panorama in order to facilitate segmentation of objects cut across by the seam. We found experimentally that using $1/8$ of the image width is a useful value for indoor scenes to ensure that objects of interest become visible in an unseparated way on at least one side.

A. Equirectangular content processing

Yolact++ uses ResNet-101 [13] with FPN [19] as its backbone. We thus replace the convolutions in the first layer of the backbone network with the equirectangular convolutions proposed in [23], leaving the parameters of the convolutions otherwise unchanged. In particular, we use the EquiConv Pytorch implementation[1]. These convolutions change which pixels are used as input depending on the position, simulating regular sampling on a spherical surface. This includes handling wrap-around, i.e., accessing pixels from the opposite image border when necessary.

While EquiConv is only used in one layer, the runtime difference in inference is still noticable, compared to the highly optimized implementations for regular convolutions, which are increasingly available (including on mobile devices).

B. Mollweide projection

The Mollweide projection [17] is a pseudocylindrical, equal-area projection. It is also known as homolographic projection or elliptical projection. In contrast to the equirectangular projection, it does not stretch areas near the poles. As a downside, the Mollweide projection bends vertical longitude lines, whereas the equirectangular projection keeps them straight. So each projection has its strong points as well as weak points. In order to retain at least to a certain degree the desirable properties of both projections, we propose a mix of both Mollweide and equirectangular projection, which we will term *hybrid Mollweide projection* in the following. We define a blending factor α in the range $[0,1]$, which allows use to interpolate smoothly between the two projections. We retrieve the standard Mollweide projection by setting $\alpha = 0.0$, the equirectangular projection by setting $\alpha = 1.0$ and a mix where both projections are weighted equally by setting $\alpha = 0.5$. The implementation of standard Mollweide and hybrid Mollweide projections follows the equations given in [26] for the equirectangular projection, with a few modifications in some places. Specifically, the equations for the conversion between the sampling point (u,v) and longitude-latitude (ϕ,θ) have to be modified properly in the following way: Equation (6) from [26] is to be replaced by

$$u = (x+0.5) \,/\, W'$$

with

$$W' = ((1-\alpha)\; d(\theta) + \alpha) \cdot W$$

$$d(\theta) = \sqrt{1 - \left(\frac{2}{\pi}\,\theta\right)^2}$$

Another point we have to take into account is that the longitude ϕ is cyclic, meaning that the image pixels on the left and right border of the Mollweide projection actually belong to the same region on the sphere. To address this, we add additional border pixels in each image row, on the left and right side. The border pixels are taken from the respective inner region of the other side (so the border pixels added on the left side are taken from the inner region of the right side, and vice versa). Figure 1 shows examples of (hybrid) Mollweide projections.

IV. Evaluation

A. Dataset preparation

For evaluation we require a dataset that provides natural panoramic images of indoor scenes. A number of the indoor datasets containing panoramic images, such as InteriorNet [18] and Structured3D [27], contain only synthetic images. Sun-CG [21] (and the derived SUMO dataset) were very actively used datasets for this purpose, but the dataset has been withdrawn. Thus there are two remaining datasets that meet this condition: Matterport3D [4] and 2D-3D-S [1]. As Matterport3D contains rather private homes than office spaces, we selected this dataset. The dataset contains 10,800 panoramic views of 90 houses. As we use a model trained on the COCO dataset, we only use the test split of Matterport3D, consisting of 18 houses with 1,848 panoramic views.

While panoramic RGB images are provided with the dataset, the instance and semantic segmentation ground truth maps are not. The scenes have been labelled on 3D meshes, and thus the annotations are provided in this format.

[1] https://github.com/palver7/EquiConvPytorch

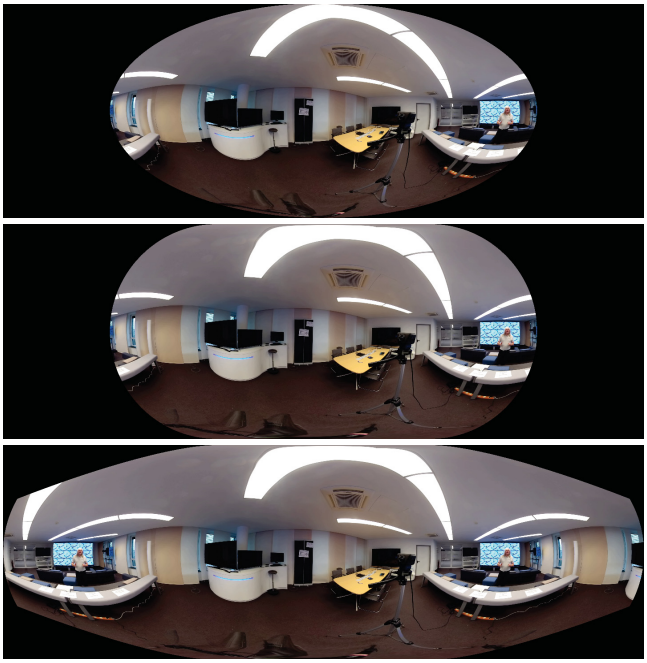

Fig. 1. Example of Mollweide projection (top), hybrid Mollweide projection ($\alpha = 0.5$, middle) and hybrid Mollweide projection with border ($\alpha = 0.5$, border=0.125, bottom).

We modified the mpview tool, that is provided with the dataset[2], in order to batch render the semantic and instance segmentation maps corresponding to each view. As support for 360° cameras is not easy to integrate into this viewer, we generate the segmentation maps for each of the 18 tiles used to compose the panoramas in the dataset, and perform the stitching process for the segmentation maps. Apart from some mislabelled parts of the mesh, some object and wall meshes have holes, that makes other objects visible (e.g., a TV screen from the outside view of a house). These issues, that cannot be resolved automatically, create some level of noise in the annotations which we have to accept due to lack of resources to manually fix them.

The annotations are provided for a set of 40 indoor classes specific for this dataset. These classes mostly (though not fully) overlap with the more commonly used NYU40 set of classes [20]. In order to work with models pretrained on the COCO dataset, we use the overlapping set of classes between Matterport3D and COCO: chair, couch, potted plant, bed, dining table, toilet, TV, sink.

Most semantic and instance segmentation methods support the COCO annotation format. We have thus created a tool to convert the Matterport3D segmentation maps to COCO annotations. This involves generating polygons from the object masks, for which we use pycococreator[3]. The COCO annotation format does not support the notion of subtracting partial polygons, thus we apply hole filling to the binary mask. In order to reduce the issue of border pixels or small regions caused by triangles cutting through the surface of other objects, we also apply morphologic closing. However, this does in many cases not remove the false annotations caused by holes in the mesh mentioned above.

One other property of the Matterport dataset is that many of the rooms are rather "loft-style", i.e., other capture locations are visible in the background. Most objects thus appear multiple times, once quite prominently in the room being captured, and one or more times in another (part of the) room. This also results in a large number of small annotated objects. In fact, 64.8% of the object instances are smaller than 0.05% of the image area, and 85.1% of the object instances occur more than once. The size differences are significant: in 62.8% of cases the smallest occurrence has an area of 1% or less than the largest occurrence of the same object instance.

For equirectangular images, annotations extending across the seam of the image will result in polygons at the left and right borders of the image. For the cases where borders for handling wrap-around have been added (either to equirectangular, Mollweide or hybrid images), we process the annotations in the border regions to keep only those that continue from the image center into the border, but remove those that only start in the border regions (and are likely to wrap around, unless they are small).

Our toolchain for preparing the Matterport3D dataset is made available at `https://github.com/atlantis-ar/matterport_utils`. It consists of a modified version of the mpview tool for generating class and instance segmentation maps, a script of combining source images and generated maps into panoramas, and a script for creating COCO style annotation files.

B. RESULTS

We compare the performance of a Yolact++ model trained on COCO applied to the panoramic Matterport3D views under different conditions in terms of projection, convolution type and wrap-around handling. We measure the average precision (AP) of detected masks at overlaps (IoU) of 50% (AP@50) and 70% (AP@70). An overview of the results is provided in Table I. In order to show the impact of the many small regions from far away objects, we provide results for evaluating against the unfiltered ground truth as well as against a ground truth where objects smaller than 0.5% of the image area have been filtered out. Note that this is a very conservative choice, that will not remove all the multiply depicted objects, but has been chosen to ensure that no smaller foreground objects are removed. To put the results in relation, it is worth noting that the current state of the art for 2D instance segmentation on the ScanNet benchmark[4] is 0.358 in terms of AP@50 (consisting of regular rather than panoramic images).

From the results we can see that there is no significant difference between using regular convolutions and equirectangular convolutions. Also the configurations adding extra

[2] `https://github.com/niessner/Matterport/tree/master/code/gaps/apps/mpview`

[3] `https://github.com/waspinator/pycococreator`

[4] `http://kaldir.vc.in.tum.de/scannet_benchmark/semantic_instance_2d.php?metric=ap50`

Projection	conv	min	wrap	border	AP	
					IoU50	**IoU70**
Equirect	regular	0.0	no	0	15.07	6.64
Equirect	regular	0.5	no	0	26.88	12.77
Equirect	equiconv	0.5	yes	1/8	26.82	12.49
Equirect	regular	0.5	no	1/8	25.21	11.86
Equirect	regular	0.5	yes	1/8	24.92	11.78
Mollweide $\alpha = 0.0$	regular	0.5	no	1/8	16.10	6.63
Mollweide $\alpha = 0.0$	regular	0.5	yes	1/8	15.97	6.58
Mollweide $\alpha = 0.5$	regular	0.5	no	1/8	21.95	10.10
Mollweide $\alpha = 0.5$	regular	0.5	yes	1/8	21.82	9.93

TABLE I

OVERVIEW OF THE RESULTS OBTAINED WITH A YOLACT++ MODEL TRAINED ON COCO. *conv* REFERS TO TYPE OF CONVOLUTION USED, *min* DESCRIBES THE MINIMUM AREA OF OBJECTS (IN PERCENT OF THE IMAGE AREA) THAT WERE RETAINED IN THE GROUND TRUTH, *wrap* DESCRIBES WHETHER WRAP AROUND HANDLING HAS BEEN APPLIED TO THE GROUND TRUTH AND *border* SPECIFIES THE WIDTH OF A BORDER BEING ADDED.

borders for wrap around handling perform very similar, though slightly worse. In addition, filtering the ground truth to have each object in the border region only once performs slightly worse (for all projections and overlaps) than not doing so. The reasons seem to be that it is not necessarily the more prominent version of the object that is better segmented, and that partial objects at borders are sometimes quite well segmented. The pure Mollweide projection performs clearly worse, and the results improve when we mix the projections.

Figure 2 shows an example of the results obtained with the different configurations and the detection of some objects, e.g., the second chair, the falsely detected TV in the office and the bed visible from the room next door. While the Mollweide projection performs generally worse than equirectangular, there are some objects that are only detected in Mollweide projection, and get lost already in the hybrid projection. We also observe some differences between the equirectangular projection with and without border. The presumption is that the aspect ratio change due to adding the border also plays a role in this behaviour.

We have performed further experiments with the SOLOv2 [24] framework, training it on the COCO and ScanNet datasets. The results indicate that the small differences in terms of performance between regular or equirectangular convolutions also hold for other models and datasets. However, the dataset determines how well the model generalises to equirectangular images. We observe that models trained on COCO images generally provide better segmentation quality (in particular, concerning the accuracy of the mask) for equirectangular images than those trained on ScanNet.

V. CONCLUSION

In this paper we have studied the problem of efficient 2D instance segmentation of 360° images of indoor scenes. We have analysed different ways of preparing equirectangular content, and assessed the use of regular vs. equirectangular convolutions on equirectangular projections. In addition, we consider the Mollweide projection as an alternative projection. We performed evaluation for the different configurations on panoramic images from the Matterport3D dataset. One contribution of this paper is thus a toolchain for preparing the panoramic dataset, and provide class and instance labels in COCO-style annotation format for use with a wide range of object detection and segmentation methods.

The conclusion from our experiments is that using equirectangular convolutions does not improve performance, but is computationally less efficient than the well optimised implementations for regular convolutions. While the Mollweide projection allows for segmentation of otherwise missed objects in a number of cases, the overall results do not outperform those on equirectangular projection. It needs to be further studied, if combining results from different projections provides benefits and justifies the increased computational effort.

REFERENCES

[1] I. Armeni, S. Sax, A. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *ArXiv*, vol. abs/1702.01105, 2017.

[2] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact++: Better real-time instance segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[3] J. Cao, R. M. Anwer, H. Cholakkal, F. S. Khan, Y. Pang, and L. Shao, "Sipmask: Spatial information preservation for fast image and video instance segmentation," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 1–18.

[4] A. Chang, A. Dai, T. Funkhouser, M. Halber, M. Niessner, M. Savva, S. Song, A. Zeng, and Y. Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *International Conference on 3D Vision (3DV)*, 2017.

[5] L.-C. Chen, A. Hermans, G. Papandreou, F. Schroff, P. Wang, and H. Adam, "Masklab: Instance segmentation by refining object detection with semantic and direction features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4013–4022.

[6] B. Coors, A. Paul Condurache, and A. Geiger, "Spherenet: Learning spherical representations for detection and classification in omnidirectional images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 518–533.

[7] A. Dai, A. X. Chang, M. Savva, M. Halber, T. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[8] M. Eder and J.-M. Frahm, "Convolutions on spherical images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 1–5.

[9] M. Eder, T. Price, T. Vu, A. Bapat, and J.-M. Frahm, "Mapped convolutions," arXiv:1906.11096, Tech. Rep., 2019.

[10] F. Engelmann, M. Bokeloh, A. Fathi, B. Leibe, and M. Nießner, "3d-mpa: Multi-proposal aggregation for 3d semantic instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9031–9040.

[11] J. Guerrero-Viu, C. Fernandez-Labrador, C. Demonceaux, and J. J. Guerrero, "Whats in my room? object recognition on indoor panoramic images," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 567–573.

[12] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2961–2969.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

Fig. 2. Example from Matterport3D dataset: input image (upper left), results on equirectangular projection with standard convolution (upper right) / equiconv (middle left), with wraparound (middle right), Mollweide projection with standard convolution (lower left: $\alpha = 0.0$, lower right: $\alpha = 0.5$). Best viewed in color.

[14] J. Hou, A. Dai, and M. Nießner, "3d-sis: 3d semantic instance segmentation of rgb-d scans," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4421–4430.

[15] Y.-C. Hsu, Z. Xu, Z. Kira, and J. Huang, "Learning to cluster for proposal-free instance segmentation," in *2018 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2018, pp. 1–8.

[16] L. Jiang, H. Zhao, S. Shi, S. Liu, C.-W. Fu, and J. Jia, "Pointgroup: Dual-set point grouping for 3d instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4867–4876.

[17] M. Kennedy and S. Kopp, *Understanding map projections*. Redlands, California: Esri Press, 2011.

[18] W. Li, S. Saeedi, J. McCormac, R. Clark, D. Tzoumanikas, Q. Ye, Y. Huang, R. Tang, and S. Leutenegger, "Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset," in *British Machine Vision Conference (BMVC)*, 2018.

[19] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.

[20] P. K. Nathan Silberman, Derek Hoiem and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[21] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1746–1754.

[22] Y.-C. Su and K. Grauman, "Learning spherical convolution for fast features from 360 imagery," in *Advances in Neural Information Processing Systems*, 2017, pp. 529–539.

[23] K. Tateno, N. Navab, and F. Tombari, "Distortion-aware convolutional filters for dense prediction in panoramic images," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 707–722.

[24] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[25] W. Yang, Y. Qian, J.-K. Kämäräinen, F. Cricri, and L. Fan, "Object detection in equirectangular panorama," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2190–2195.

[26] Y. Ye, E. Alshina, and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360lib," Joint Video Exploration Team (JVET), Tech. Rep., 2017.

[27] J. Zheng, J. Zhang, J. Li, R. Tang, S. Gao, and Z. Zhou, "Structured3d: A large photo-realistic dataset for structured 3d modeling," in *Computer Vision – ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 519–535.