What’s in my Room? Object Recognition on Indoor Panoramic Images

Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, Jose J. Guerrero

What's in my Room? Object Recognition on Indoor Panoramic Images

Julia Guerrero-Viu*[1], Clara Fernandez-Labrador*[1,2], Cédric Demonceaux[2], and Jose J. Guerrero[1]

Abstract— **In the last few years, there has been a growing interest in taking advantage of the 360° panoramic images potential, while managing the new challenges they imply. While several tasks have been improved thanks to the contextual information these images offer, object recognition in indoor scenes still remains a challenging problem that has not been deeply investigated. This paper provides an object recognition system that performs object detection and semantic segmentation tasks by using a deep learning model adapted to match the nature of equirectangular images. From these results, instance segmentation masks are recovered, refined and transformed into 3D bounding boxes that are placed into the 3D model of the room. Quantitative and qualitative results support that our method outperforms the state of the art by a large margin and show a complete understanding of the main objects in indoor scenes.**

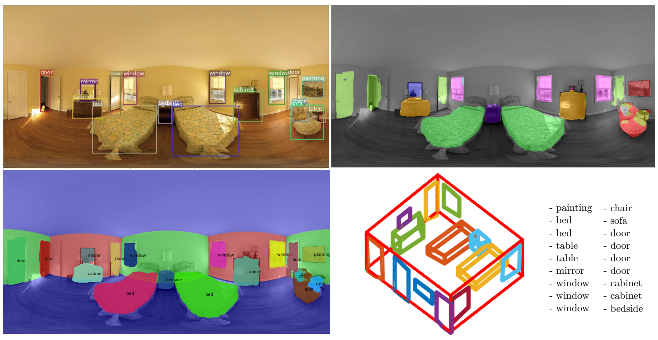

Fig. 1: **What's in my room?**: from a single spherical RGB image, we obtain localization and pixel-wise classification of the objects present in the scene. We use this information to recover instance segmentation masks of all objects and finally place them inside the 3D room layout.

I. Introduction

The increasing interest in autonomous mobile systems, like drones, robotic vacuum cleaners or assistant robots, makes detection and recognition of objects in indoor environments a very important and demanded task.

Since recognizing a visual concept is relatively trivial for a human, it is worth considering the hard challenges inherently involved. Objects in images can be oriented in many different ways, vary their size, be occluded, blended into the environment because of their color or appearance, or affected by different illumination conditions, which changes drastically their aspect on the pixel level. Moreover, the concept behind an object's name is sometimes broad, including non-clear frontiers to other concepts. For example, where do you consider the limits between a sofa and an armchair?

Convolutional Neural Networks (CNNs) have already demonstrated to be the best known models to perform object recognition, as they are capable of dealing with those challenges by automatically learning objects' inherent features and correctly identify their intrinsic concepts.

However, images from conventional cameras have a small field of view, much smaller than human vision, which implies that contextual information cannot be as useful as it should. To overcome this limitation, a real impact came with the arrival of the 360° full-view panoramic images, which are recently arising more and more interest in the robotics and computer vision community, as they allow us to visualize, in a single image, the whole scene at the same time. Together with all of their potential we have to deal with challenges produced by their own spherical projection, such as distortion, or the lack of complete, labeled and massive datasets. This requires the development of specific techniques that take advantage of their strengths and allow working with panoramic images in an efficient and effective way.

In this paper, we propose an object recognition system that provides a complete understanding of the main objects in an indoor scene from a single 360° image in equirectangular projection. Our method extends the BlitzNet model [1] to perform both object detection and semantic segmentation tasks but adapted to match the nature of the equirectangular image input. We train the network to predict 14 different classes of main indoor scenes related objects. Results of the CNN are post-processed to obtain instance segmentation masks, which are successfully refined by taking advantage of the spatial contextual clues that the room layout provides. In this work, we not only show the potential of exploiting the 2D room layout to improve the instance segmentation mask, but also the possibility of leveraging the 3D layout to generate 3D object bounding boxes directly from the improved masks.

II. Previous Work

Object detection field has been mainly dominated by two different approaches: one-stage and two-stage detectors. Two-stage detectors, as the first R-CNN[2] architecture followed by its variants Fast R-CNN[3], Faster R-CNN[4] and Mask R-CNN[5] achieve great accuracy but lower speed. They require firstly to refine proposals to obtain the features needed to classify the objects. On the other hand, one-stage detectors, following YOLO[6] and SSD[7] simultaneous bounding box refinement and classification,

* Equal contribution
[1]J. Guerrero-Viu, C. Fernandez-Labrador and J.J. Guerrero. Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Spain
[2]C. Fernandez-Labrador and Cédric Demonceaux. VIBOT ERL CNRS 6000, ImViA, Université Bourgogne Franche-Comté, France.

significantly reduce computational cost. They achieve real-time performing maintaining high accuracy, which is needed for most applications in autonomous mobile systems. SSD multi-scale pyramid idea proves to help in conducting more accurate detections and manage widely various object sizes, approach followed in most state-of-the-art object detectors.

While all those models optimize bounding box detection, not so many integrate in their pipeline the pixel-wise recognition needed for many applications. In this way, BlitzNet [1] is a one-stage multi-scale model that adds semantic segmentation and therefore recognizes objects at pixel level. It also proves the advantages of jointly learning two scene understanding tasks: object detection and semantic segmentation, which benefit from each other by sharing almost the complete network architecture.

However, state-of-the-art research mainly focuses on using conventional images. Their limited field of view prevents contextual information from being as crucial as it is in scene understanding for humans. Differently from outdoor object recognition, where thanks to the increasing research on autonomous driving, there are recent works using panoramic images [8] [9], there is no wide research on object recognition from indoor panoramas. The most relevant work that addresses this problem is PanoContext [10]. It includes 2D object detection and semantic segmentation among other 3D scene understanding tasks, proving the potential of having a larger field of view for recognition problems. Their method, nevertheless, is based on geometrical reasoning and traditional computer vision feature extractors and can be still considered as state-of-the-art in indoor object recognition on panoramic images. Recent research on this kind of images includes 3D layout recovery [11] [12] [13] [14] and scene modeling [15], which provides global context and gives a 3D interpretation of the scene from a single view. In [16], they show that this tasks can also benefit and augment an omnidirectional SLAM. Combining object recognition and 3D layout recovery motivates our proposal to obtain the 3D recognition and location of main objects in our room.

III. Dataset Extension

Panoramic images datasets with object recognition labels are not as standard or complete as conventional images ones [17] [18] [19]. Therefore, in this work we decide to extend the SUN360 database [17] with segmentation labels. For every panorama, we generate individual masks encoding each object's spatial layout. Additionally, we combine all the masks obtaining a semantic segmentation panoramic image with per-pixel classification without differentiating instances.

Bedroom and living room sets, formed by 418 and 248 images respectively, are used and 14 different object classes are considered. The dataset is divided into 85% for train and validation and 15% for test.

We generate segmentation masks based on 2D bounding points of the objects, taken from PanoContext [10] work. We project them on the spherical domain to follow distortion patterns in contours and to correctly manage objects that appear cropped on the horizontal image limits. To combine the binary masks and create the semantic segmentation panorama, with the lack of depth or other 3D information, an hypothesis of occlusion among objects is needed. We consider the assumption that objects are not in general completely occluded, and therefore for each pair of objects in conflict their area of overlap and size are computed. If area of overlap is bigger than a threshold, the smallest object is considered closer and completely visible and otherwise the biggest one is selected. With its evident limitations, this hypothesis experimentally proves to work well in most of the cases, allowing to correctly segment most of the visible and cuboid-shaped objects in images as shown in Figure 2.

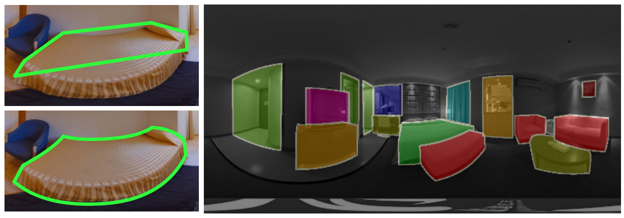

Fig. 2: Result of our method to **create semantic segmentation masks**, assuming hypothesis of occlusion. Notice on the left the differences between creating straight contours on image domain (top) vs. spherical domain (bottom).

The complete **dataset** used in this work is released for public access and can be found in the project webpage[1].

IV. Model

In this section we present our object recognition model, called *Panoramic BlitzNet*, that is based on the original CNN BlitzNet [1] but adapted to work specifically with complete equirectangular images. It addresses both object detection and semantic segmentation tasks, following BlitzNet architecture: a Fully Convolutional model that follows the encoder-decoder approach with skip connections. It performs multi-scale recognition and takes advantage of joint learning. Main changes to their base implementation include the use of the complete rectangular panorama, modifying the input aspect ratio. We also change the anchor boxes proposals, as the new input shape needs to be considered because they are centered on pixels grid. Our bounding boxes proposals are done by firstly converting image to a regular grid, covering the whole rectangular-shaped image. Grid has different dimensions in each layer, from 128x256 to 1x2, because of the iteratively lower scale of the feature maps. In each grid cell 5 different proposals are created with 5 different aspect ratios: 1, 2, 1/2, 3 and 1/3, allowing the network to manage different object shapes.

Special mention deserves data augmentation as an important technique to avoid overfitting, particularly on non-massive datasets like in our case. Here, we modify the original data augmentation by removing random crops on images (contextual information is important) and adding horizontal rotation from $0°$ to $360°$ to cover all different positions on the sphere.

[1]Available at https://webdiis.unizar.es/~jguerrer/room_OR/

A. How can we deal with 360° images distortion?

We exploit the potential of omnidirectional images covering 360° horizontal and 180° vertical field of view represented in equirectangular projection. While these images allow us to analyse the whole scene at once taking advantage of all the context, they present great distortions due to their projection of the sphere. Here, we study the impact of replacing standard convolutions with equirectangular convolutions (EquiConvs [14]) on the task of recognizing objects. With this kind of convolutions, the kernel **adapts its shape and size accordingly to the distortions** produced by the equirectangular projection. As mentioned in [14], the distortion presented is location dependent, specifically, it depends on the polar angle. They demonstrate how EquiConvs can be really convenient to generalize to different camera positions since the layout shape can suffer from many variations. For the specific task of object recognition, the use of EquiConvs is definitely convenient even if the camera is always at the same place, since objects can be at many different locations inside the scene –*e.g.* objects closer to the camera will have greater distortions than objects around the horizon line. EquiConvs here play an important role since they can learn ignoring this distortion patterns and thus, being more able to learn real objects appearance. Additionally, one important challenge to accomplish our goal is represented by the need of extensive annotations for training object recognition. To this end, EquiConvs make the **pre-training much more effective** since this type of convolutions implicitly handle equirectangular distortion, being able to use previous weights from conventional images as if they were learnt on the same kind of images. We can therefore exploit the wealth of publicly available perspective datasets for training – SUN RGB-D [20] dataset in this case, which reduces the cost of annotations and allows training under a larger variety of scenarios. Moreover, standard convolutions do not understand that the image wraps around the sphere, loosing the continuity of the scene, while EquiConvs, working directly on the spherical space, avoid padding and exploit this idea. This makes them also very suitable for the objects that appear cut between the left and right side of the image.

B. From semantic to instance segmentation

Semantic segmentation masks allow us to pixel-wise classify scenes in object categories. One step further goes instance segmentation, which classifies each pixel not only to its category but also differentiating its concrete object instance, an essential stage to correctly locate them into the 3D reconstruction of the room. Without using any instance segmentation prior to be learnt, we add a simple post-processing to obtain **pixel-wise instance classification** of the scene, based on the outputs of the network. Considering each bounding box detection as a Gaussian distribution, it is assumed that 99% of the object is contained on it, so standard deviation in each dimension is taken as $\sigma_w = \frac{width}{6}, \sigma_h = \frac{height}{6}$, giving 3σ at each side of the mean, which is defined as the center of the bounding box. Then, each pixel from the semantic segmentation mask is assigned to the instance distribution with the minimum Mahalanobis distance. When none of the distances to distributions exceeds the chi-squared test threshold, the pixel is left as the classification in the initial semantic segmentation, a proposal that shown the best experimental results.

This assumption gives, consequently, more confidence to bigger objects, which are usually better segmented by our model. In addition to providing an approach to create instance segmentation masks, we analyze the impact it can have in improving our initial segmentation, as shown in experiments.

C. Can we convert instance segmentation masks into 3D bounding boxes?

If there is a task that has experimented a disruptive innovation with the emergence of 360° images, it has been the room layout estimation problem [12], [13], [21], [14], [22]. In these works, from a single panorama, they recover the main structure of the room –*i.e.* disposition of the walls, ceiling and floor, not only in the image domain, but also a complete 3D reconstruction model up to scale.

The intuition is that objects location and pose inside a 3D indoor space are not randomly distributed. Following the law of physics, objects will be fairly constrained to lie on at least one supporting plane, in stable configurations and, in several cases, aligned with the room walls. This means that the room layout provides strong spatial contextual cues as to where and how objects can be found.

Thus, in order to provide a greater understanding of the scene, here we analyze the potential of using the room layout as a *prior* combined with the object recognition task:

1) We found that we can easily leverage the room layout in the image domain to **improve the instance segmentation masks**. Based on the contextual information given by the layout, there are a series of logical assumptions that we can immediately make –*e.g.* it is very unlikely to find doors not resting on the floor or paintings hanging in between two walls. During this process we also detect holes in masks and fill them.

2) The aforementioned methods provide 3D layout models of the rooms. This allows us to **place the identified objects inside the 3D model of the room** as long as they lie on the walls, or rest on the floor / ceiling aligned with the walls. In this way, only by detecting the masks of the objects and with a good layout prior, we can obtain a very precise 2D representation of the objects and an initial estimate of the 3D understanding of the scene.

To obtain the room layout, here we choose the work of Fernandez *et al.* [14]. Additionally, we will need to assume Manhattan World [23] whereby there exist three dominant orthogonal directions defining the scene. Here, we compute the vanishing points of the scene following the approach of [21]. They propose a RANSAC-based algorithm that works directly on omnidirectional images running up to 5 times faster than other approaches.

For the 3D object recognition task, in PanoContext [10] they generate many cuboid hypotheses combining two ap-

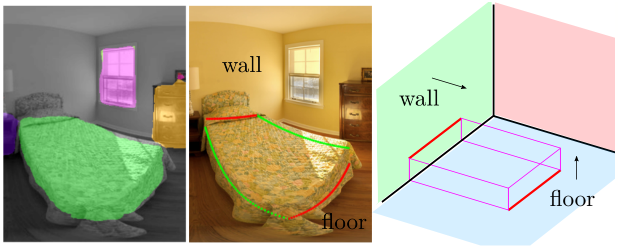

Fig. 3: **From mask to 3D**: We find the lines that best fit the object mask by a RANSAC algorithm and orient them accordingly to the main directions of the scene. The object dimensions are obtained trusting in the line resting on the wall and the line resting on the floor (red lines in the figure).

proaches. First, they perform rectangle detection in six axis-aligned views projected from the original panorama. Then, they sample rays from the vanishing points to fit image segmentation boundaries obtained by selective search. Finally, they choose the best cuboid whose projection has the largest intersection over union score with the segment. Here instead, we directly approximate every object mask with four lines by a RANSAC approach and classify these lines accordingly to the vanishing points, vp, –*i.e.* we obtain the orientation of the objects inside the scene. We consider that a line belongs to one concrete direction, $k = x, y, z$, when its normal direction in the 3D space n_i fulfills the condition of perpendicularity with the direction under an angular threshold of $\pm 0.5°$ experimentally obtained, $|arccos(n_i \cdot vp_k) - \frac{\pi}{2}| \leq \theta_{th}$. Then, we determine with which planes of the room each object interacts by obtaining the intersections between the object mask and the 2D room layout segmentation map. The room layout segmentation map is directly generated from the predicted layout corners [14], encoding each plane accordingly to the main directions of the scene. If we predict that the object lies on a wall –*i.e.* doors, windows, mirrors, pictures, etc., we simply project the object mask to its position in the wall in the 3D room layout. For cuboid objects like beds, sofas or bedside tables, we can obtain the object dimensions –*i.e.* length, width and height, using the lines interacting with the room planes. Figure 3 shows how we deal with these objects. With these ideas, we can place most of the objects inside the 3D scene and obtain a good understanding of the scene from the 2D segmentation masks. Some qualitative results are shown in Figure 4 as well as in the teaser image, Figure 1.

V. Experiments

We evaluate our model by different experiments conducted on SUN360 [17] extended dataset, which are presented in this section. Experimental setup is explained in order to make our work reproducible, together with the detailed evaluation metrics.

Experimental setup: The whole model is coded in Python 3.5 using the framework Tensorflow v1.13.1. All experiments were conducted on a single Nvidia GeForce GTX 1080 GPU.

As in [1], we use ResNet-50 as feature extractor, Adam stochastic algorithm [24] for optimization and learning rate

	TRAIN mAP^w	TEST mAP^w	TRAIN $meanIoU$	TEST $meanIoU$
From scratch	0.911	0.468	0.872	0.419
ImageNet (ResNet)	0.896	0.479	0.788	0.432
SUN RGB-D	0.728	**0.516**	0.742	**0.461**

TABLE I: **Effect of initialization:** Results tested on original BlitzNet with different weights initializations. Notice how initializing with SUN RGB-D weights gives clearly better test results. Evaluation on train set is shown to observe the overfitting effect, specially clear in trained from scratch model.

	mAP^w	mAP	$meanIoU$
BlitzNet	0.516	0.688	0.461
Panoramic BlitzNet	**0.632**	**0.768**	**0.530**

TABLE II: **Effect of adapting the CNN for panoramas:** Comparison between results on panoramic images with BlitzNet vs. our proposed *Panoramic BlitzNet*.

set to 10^{-4} and decreased twice during training. Experiments were conducted by changing that learning rate without noticeable influence. We use stride 4 in the last layer of the upscaling stream and varying mini-batch sizes, which are stated in each experiment. All models are trained until convergence, measured with a random validation subset.

Based on our dataset characteristics, we decide to pre-train our network instead of initializing it randomly, that would conduct to a clear overfit to our data, as studied in the first experiment. Because of the lack of massive panoramic datasets, conventional images are used for pre-training: Firstly, we use the publicly available weights of ResNet50 backbone on ImageNet [25] dataset. With that initialization, we then train the whole network on SUN RGB-D [20], pre-processed to have the same common classes. This way we have an initialization for the complete model to be able to fine-tune with the panoramic images, possible thanks to a Fully Convolutional network where weights can be shared with variable input dimensions.

Evaluation metrics: To evaluate detection performance we use typical mean average precision (mAP), considering that a predicted bounding box is correct if its intersection over union with the ground truth is higher than 0.3, as exact localization is better predicted in the segmentation branch. The average precision evaluates interpolated precision at all different recall levels, in its simplest definition, but can also be widely found in literature as weighted by the recall area that they represent. In this paper we mostly use simple AP and when using the second version it is referred as AP^w. Finally, when calculating the mean among all different classes we use a weighted mean defined as follows,

$$mAP = \frac{1}{M} \sum_{i=1}^{M} \frac{d_i}{n} AP_i \tag{1}$$

being M the number of classes, AP_i the average precision per class, d_i the number of detections of class i and n the total number of detections. It is considered as a more representative metric because results calculated from objects with a minimum number of samples in the test set should be less significant when analyzing a global performance.

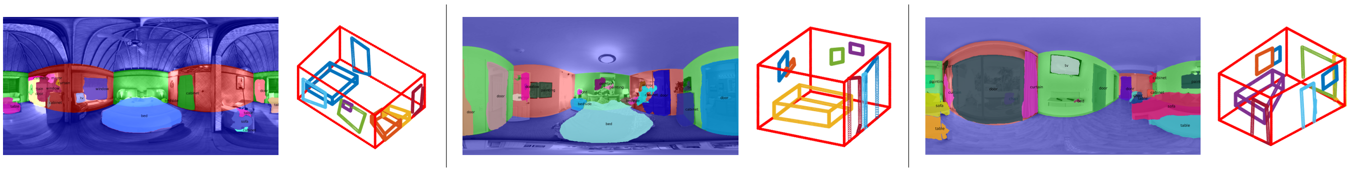

Fig. 4: Examples of 3D models obtained from instance objects masks and room layout knowledge [14].

input	mIoU	backgrd.	bed	painting	table	mirror	window	curtain	chair	light	sofa	door	cabinet	bedside	tv	shelf
Initial seg. map	53.0	90.7	61.7	**32.1**	**75.2**	**42.3**	55.8	54.0	55.1	31.4	**34.6**	**63.6**	48.5	**40.7**	52.2	57.4
Post-processed seg. map	**53.1**	90.7	**63.3**	30.1	75.0	41.5	**56.7**	**55.3**	**55.5**	**34.0**	31.8	62.9	**48.8**	40.2	**53.5**	**57.5**

TABLE III: Semantic segmentation results before and after applying the **instance segmentation post-processing**. Initial semantic segmentation is taken from our CNN output.

Segmentation performance is measured with mean intersection over union ($mIoU$) between predicted and ground truth segmentation maps.

A. Initialization

First experiment was conducted before the development of our model, to verify the importance of pre-training. It uses original BlitzNet300 architecture, without modifying the base implementation to adapt for panoramas. Batch size is set to 16 and it is trained until convergence. When training, three different initializations are executed to compare: random initialization (trained all from scratch), ImageNet initialization of the feature extractor (ResNet) and pre-trained SUN RGB-D initialization of the complete model. Results are compared in Table I, which shows that pre-training with SUN RGB-D followed by fine-tuning the complete network with panoramic dataset, gives the best performance results and generalizes better. ImageNet initialization converges faster, but it demonstrates higher overfitting than SUN RGB-D pre-training. This experiment shows that given a relatively small panoramic dataset, the use of a massive one of conventional images for pre-training allows the network to learn higher level characteristics of the objects, avoiding overfitting and being one of the keys for the success of the system.

B. Square vs. Panoramic

This experiment compares the performance of our *Panoramic BlitzNet* network with the original model designed for conventional images. In this case, batch size is reduced to 4, for memory limitations in our GPU. As seen in Table II, our adapted model demonstrates clearly better results, improving performance by a wide margin and supporting our assumption of the important benefits of a concrete model designed for panoramas.

Apart from avoiding distortions and crops to make it fit to a square shape, it also shows the benefits of using a wider field of view, which allows to consider the whole context of the room. This idea is a strong support for the potential of panoramic images, not only in object recognition but in many other visual tasks, at least in those related to indoor scene understanding problem.

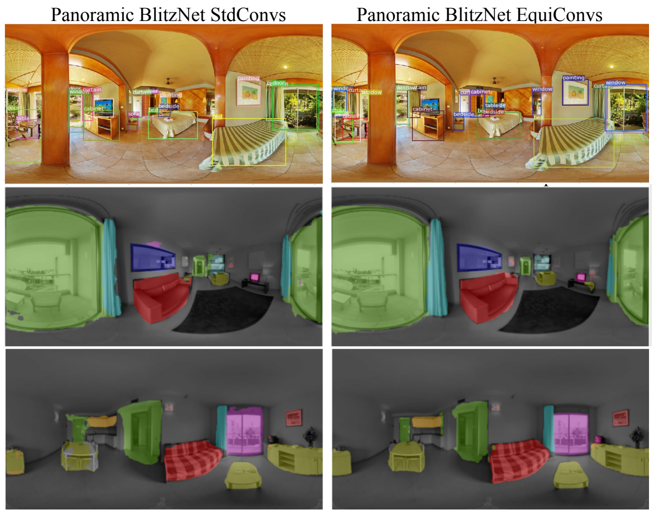

Fig. 5: **Qualitative evaluation of object detection and semantic segmentation:** Examples of results obtained with our *Panoramic BlitzNet* using both standard convolutions and EquiConvs [14].

C. StandardConvs vs. EquiConvs

Evaluation of the influence that equirectangular convolutions have on object recognition task is a key point in this work. Comparison of both models can be seen in Tables IV and V and Figure 5. There we show that our model with EquiConvs, adapting the kernel to manage the equirectangular distortion, perform better in both object detection and semantic segmentation tasks. Additionally, equirectangular convolutions have other advantages over standard convolutions. They make our pre-training on conventional images more meaningful, as managing distortions by the kernel allows to share the weights as if they were trained on the same kind of images. Therefore we strongly believe that this type of convolutions help in avoiding overfitting to training data, which due to the particularities of the SUN360 dataset (camera pose does not vary and scenes are relatively similar) help test results, and will probably be crucial when working on different datasets. Finally, EquiConvs also prove to make detections with higher confidence as, when raising the confidence threshold to 0.95, their recall is maintained

model	mAP	bed	painting	table	mirror	window	curtain	chair	light	sofa	door	cabinet	bedside	tv	shelf
$*$ DPM	29.4	35.2	56.0	21.6	19.2	21.8	29.5	26.0	—	22.2	31.9	—	—	31.0	—
$*$ Deng *et al.*	68.7	76.3	68.0	73.6	58.7	62.6	69.5	68.0	—	72.5	67.3	—	—	70.0	—
PanoBlitzNet StdConvs	76.8	94.9	**85.0**	**83.3**	71.9	**72.2**	72.2	71.9	35.0	**89.3**	**75.5**	**57.9**	87.9	91.1	30.5
PanoBlitzNet EquiConvs	**77.8**	**95.3**	83.9	82.1	**76.2**	70.9	**75.9**	**80.9**	**41.0**	85.4	72.5	55.6	**91.4**	**93.3**	**40.2**

TABLE IV: **Object detection** results on SUN360 test set with our method *Panoramic BlitzNet* using different convolutions (Standard vs. Equirectangular). $*$ Results trained and evaluated on a combination of datasets (including SUN360) by Deng *et al.* [26]

model	mIoU	backgrd.	bed	painting	table	mirror	window	curtain	chair	light	sofa	door	cabinet	bedside	tv	shelf
PanoContext	37.5	86.9	**78.6**	38.7	29.6	38.2	35.6	—	09.6	—	11.1	19.4	27.4	39.7	34.8	—
PanoBlitzNet StdConvs	53.0	90.7	61.7	32.1	**75.2**	**42.3**	**55.8**	**54.0**	55.1	**31.4**	**34.6**	63.6	48.5	**40.7**	52.2	57.4
PanoBlitzNet EquiConvs	**54.4**	**91.3**	62.1	**61.2**	72.3	41.1	53.4	53.7	**55.2**	26.5	32.9	**63.8**	**51.1**	36.6	**52.3**	**61.9**

TABLE V: **Semantic segmentation** results on SUN360 extended test set with our proposed model *Panoramic BlitzNet* using different convolutions (Standard vs. Equirectangular) and comparison with PanoContext [10].

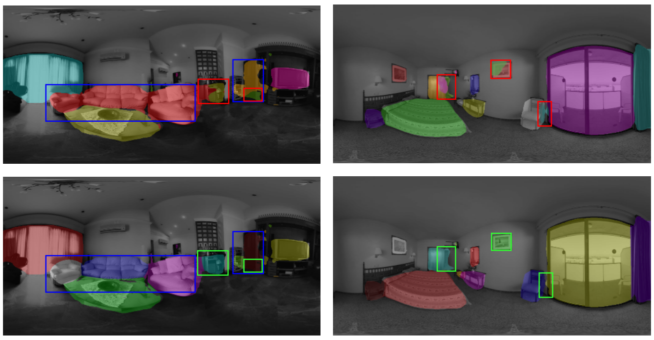

Fig. 6: **Instance segmentation post-processing** results. Top is initial semantic segmentation (output of CNN) and bottom is result of post-processing. Notice that apart from correctly differentiate among instances (highlighted in blue) it improves original segmentation (highlighted in red and green for failed and improved segmentation respectively).

over 40% compared to 28% achieved with standard convolutions. Further qualitative results can be found in our project webpage [1].

D. Instance segmentation post-processing

In this experiment we compare the semantic segmentation output of the network with the result of applying our instance segmentation method to create improved semantic segmentation maps. Although this is not the objective of the post-processing (the goal is to differentiate among different instances), it is evaluated to prove the influence that it can have in segmentation performance. Our intuition was that the instance segmentation method would imply an improvement to the initial segmentation maps because it gives higher confidence to detections, whose performance is clearly higher than segmentation's one in our model. Results, shown in Table III, are very similar and lead us to conclude that the post-processing does not prove to be influential in this way. However, qualitative results support our intuitive idea by showing some clearly improving cases that are remarked in Figure 6.

Fig. 7: After combining the room layout with our segmentation masks, the model experiences a clear improvement as a whole. However, here we want to show a **failure case** where, when assuming that doors must reach the floor, we may have overlapping with other occluding objects in the image, damaging segmentation results but improving the door 3D localization.

Apart from that, instance segmentation maps are qualitatively evaluated as seen in Figure 6. Our approach proves to work well on several different scenes by correctly separating same category objects, that initially overlapped in semantic maps, into different instances. Limitations of the method can be seen when the network fails detecting an object, which is therefore not differentiated as an instance on the final map and when managing objects with complex shapes that can not be modelled with a gaussian distribution.

Here, we finally analyze the improvement over our segmentation masks by leveraging the contextual information of the room layout. In our experiment, logical assumptions used for this refinement entail a significant improvement of up to 7.2% mIoU with respect to the segmentation output of *Panoramic BlitzNet* with StdConvs, achieving a final **mIoU = 60.3%**. It should be noted that the classes that contribute most to this improvement are mirror, window and painting. However, while one would also expect a clear improvement in the door category, we have seen a drop in performance in some cases such as the one shown in Figure 7, although it definitely has a positive effect on its location in the 3D room space. As already supported by this preliminary experiment, we propose a promising method to noticeably benefit 2D and 3D object recognition tasks from room layout knowledge, and encourage the idea that it is worth continuing to work in this direction.

E. Comparison with the State of the Art

Detection: In Table IV we show our detection results on the SUN360 extended dataset. Our Panoramic BliztNet with EquiConvs achieves very satisfactory results, with a global **mAP = 77.8%**. Here we include the results of [26], recent work on indoor panoramic object recognition, where Deng *et. al* also evaluated their own implementation of DPM [27] tested on panoramas. Our method achieves the best results in detection for all 10 common classes compared to them. It is worth noting that our approach achieves these results just training with ~ 400 panoramas from the SUN360 dataset while they use additional panoramas to train their model. Since their dataset is not public and no code is available, we report directly the results collected in [26].

Segmentation: Table V summarizes the semantic segmentation results on the SUN360 extended dataset. A direct comparison is possible with the work of PanoContext [10]. The results clearly show that our method significantly improves over the state of the art. In particular, we add three new object classes and boost **mIoU = 54.4%**, which represents an improvement of 16.9% over PanoContext's method.

VI. Conclusion

From a single panoramic image, we propose a method that provides a complete understanding of the main objects in an indoor scene. By managing the inherent characteristics and challenges that equirectangular panoramas involve, we outperform state of the art in addition to creating a more complete system, which not only obtains 2D detection and pixel-wise segmentation of objects but also places them into a 3D reconstruction of the room. Exploiting the advantages of having a wider field of view in indoor environments, this visual system becomes a promising key element for future autonomous mobile robots. Future work includes the inclusion of instance segmentation predictions into the deep learning pipeline and a further study of the potential in combining layout recovery and object recognition tasks.

Acknowledgment

Work supported by Project RTI2018-096903-B-I00 (AEI/FEDER, UE) and Regional Council of Bourgogne Franche-Comte (2017-9201AAO048S01342).

References

[1] N. Dvornik, K. Shmelkov, J. Mairal, and C. Schmid, "Blitznet: A real-time deep network for scene understanding," in *ICCV*, 2017, pp. 4154–4162.

[2] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[3] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[5] 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.

[6] J. Redmon and A. Farhadi, "Yolo9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 7263–7271.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.

[8] X. Meng, X. Zhang, K. Yan, and H. Zhang, "Real-time detection and recognition of live panoramic traffic signs based on deep learning," in *2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2018, pp. 584–588.

[9] 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.

[10] Y. Zhang, S. Song, P. Tan, and J. Xiao, "Panocontext: A whole-room 3d context model for panoramic scene understanding," in *European conference on computer vision*. Springer, 2014, pp. 668–686.

[11] J. Omedes, G. López-Nicolás, and J. J. Guerrero, "Omnidirectional vision for indoor spatial layout recovery," in *Robotics and Autonomous Systems*, 2014, pp. 62(9):1271–1281.

[12] H. Jia and S. Li, "Estimating structure of indoor scene from a single full-view image," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 4851–4858.

[13] C. Zou, A. Colburn, Q. Shan, and D. Hoiem, "Layoutnet: Reconstructing the 3d room layout from a single rgb image," in *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2051–2059.

[14] C. Fernandez-Labrador, J. M. Facil, A. Perez-Yus, C. Demonceaux, J. Civera, and J. J. Guerrero, "Corners for layout: End-to-end layout recovery from 360 images," *arXiv preprint arXiv:1903.08094*, 2019.

[15] Y. Yang, S. Jin, R. Liu, S. Bing Kang, and J. Yu, "Automatic 3d indoor scene modeling from single panorama," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[16] R. Lukierski, S. Leutenegger, and A. J. Davison, "Room layout estimation from rapid omnidirectional exploration," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 6315–6322.

[17] A. O. J. Xiao, K. A. Ehinger and A. Torralba, "Recognizing scene viewpoint using panoramic place representation," *Proceedings of 25th IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[18] S. Song, F. Yu, A. Zeng, A. X. Chang, M. Savva, and T. Funkhouser, "Semantic scene completion from a single depth image," *Proceedings of 30th IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[19] I. Armeni, S. Sax, A. R. Zamir, and S. Savarese, "Joint 2d-3d-semantic data for indoor scene understanding," *arXiv preprint arXiv:1702.01105*, 2017.

[20] S. L. S. Song and J. Xiao., "SUN RGB-D: A RGB-D Scene Understanding Benchmark Suite," *Proceedings of 28th IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[21] C. Fernandez-Labrador, A. Perez-Yus, G. Lopez-Nicolas, and J. J. Guerrero, "Layouts from panoramic images with geometry and deep learning," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3153–3160, 2018.

[22] C. Sun, C.-W. Hsiao, M. Sun, and H.-T. Chen, "Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1047–1056.

[23] J. M. Coughlan and A. L. Yuille, "Manhattan world: Compass direction from a single image by bayesian inference," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2. IEEE, 1999, pp. 941–947.

[24] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[25] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[26] F. Deng, X. Zhu, and J. Ren, "Object detection on panoramic images based on deep learning," in *2017 3rd International Conference on Control, Automation and Robotics (ICCAR)*. IEEE, 2017, pp. 375–380.

[27] D. A. Forsyth, "Object detection with discriminatively trained part-based models," *IEEE Computer*, vol. 47, pp. 6–7, 2014.