22^{nd} *Computer Vision Winter Workshop*
Nicole M. Artner, Ines Janusch, Walter G. Kropatsch (eds.)
Retz, Austria, February 6–8, 2017

Semantic segmentation for room categorization

Ajda Lampe, Luka Čehovin, Rok Mandeljc, Matej Kristan
Faculty of computer and information science, University of Ljubljana
Večna pot 113, Ljubljana, Slovenia
ajda.lampe@gmail.com {luka.cehovin,rok.mandeljc,matej.kristan}@fri.uni-lj.si

Abstract. *Room category recognition is a highly desired functionality in household robots as it enables natural indoor navigation without having to rely solely on range sensors. Significant research has been invested into the problem of image-based room recognition. All approaches assume that the image contains a single category and usually apply holistic scene models. This assumption is often violated in realistic robotic scenarios, in which a robot observes several room categories within its view. We propose an alternative approach for room category recognition, based on per-pixel semantic segmentation. Our approach decomposes the input image into semantic regions, and easily deals with images containing either a single or multiple categories. As a side product, each room category is localized in the image, which makes our approach suitable for robotic exploration and navigation based on visual feedbacks. Our approach is evaluated on a new dataset, annotated with pixel-wise ground-truth labels for eight room categories. Results show that our approach outperforms the state-of-the-art in recognition as well as pixel-wise localization. Our approach exhibits a remarkable robustness to rotation and outperforms the holistic state-of-the-art approach even on examples with a single category.*

1. Introduction

Visual room categorization is a high-level computer vision task in which the system is presented with an image of an indoor scene, and is expected to predict the most likely semantic label of the room depicted in the image. Such systems are extremely useful in mobile robotics, where the maps of environment, used for navigation, do not include such semantic information. A typical household robot would need to recognize the semantic meaning of its current location in order to successfully reason about the given tasks, and to be able to navigate the surroundings and perform operations in appropriate rooms.

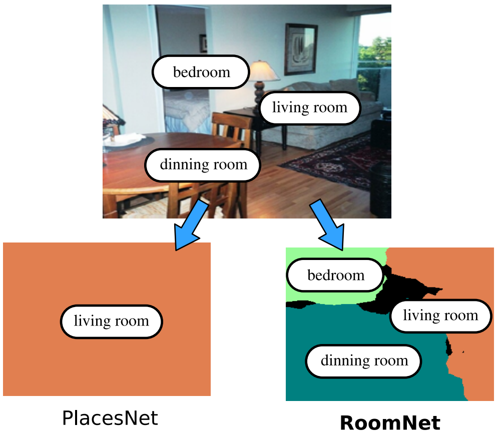


Figure 1: Example of an image with multiple room categories present. Holistic approaches, such as PlacesNet [23], can only report a single (dominant) room category, while our approach produces a coarse segmentation that enables multi-category recognition as well as in-image category localization.

Recently, computer-vision-based room classification methods have achieved high classification accuracy by employing convolutional neural networks, trained on large annotated datasets of indoor and outdoor scenes [23, 6, 21]. However, the main limitation of these approaches lies in the formalization of the classification task, where the image is classified *holistically*, i.e., the entire scene is assigned a single semantic label. This is true even for methods that otherwise utilize spatial information [9, 13]. In realistic scenarios, the observed indoor scenes frequently contain multiple room categories. Figure 1 shows an example with different rooms and areas with distinct semantic functionality present in a single view. In such cases, the majority of existing room classification approaches is limited to producing only a single output label. Some of the approaches do provide a probability distribution over the possible room categories, but this is still insufficient for accurate category localization within the robot field of view.

We propose a change of paradigm in room categorization by casting it as a problem of approximate semantic segmentation. Our approach uses local and contextual information to infer the most likely room category at each pixel. The final room category is estimated by aggregating per-pixel classifications. In contrast to existing approaches that apply holistic aggregation into a single category, our approach detects clusters of category labels in the image and can be considered as non-parametric estimation. Thus the approach is expected to be more robust than related approaches in cases when observing an image with a single as well as multiple categories. As a side product, our approach allows category localization within an image.

We claim two contributions. The first contribution is a novel approach for room classification and localization, based on semantic segmentation, which we call the RoomNet. To the best of our knowledge, this is the first method that addresses the problem of room classification through semantic segmentation. Our second contribution is a new dataset for room classification. In contrast to existing datasets, our dataset contains images with a single and multiple categories and is per-pixel labeled with the category identities. Our RoomNet is evaluated extensively on this dataset. Results conclusively show that the RoomNet outperforms a recent state-of-the-art room classification approach by approximately 15% on multiple as well as single category images, is more robust to view rotation and performs accurate localization.

The remainder of this paper is organized as follows. Section 2 overviews the related work, Section 3 describes the proposed room categorization approach, Section 4 presents the experimental evaluation, and conclusions are drawn in Section 5.

2. Related work

Majority of existing literature in the field of vision-based room categorization employs holistic descriptors; they construct a feature representation corresponding to the entire image, and classify it with a multi-class classifier. The traditional approaches rely on bag-of-words representations, constructed from low-level features. In [3], authors construct a "bag of keypoints" representation from keypoints, detected using Harris affine detector and encoded by SIFT descriptor; the resulting feature vector is classified using naive Bayes classifier. In [8], the holistic bag-of-words descriptor is constructed from HOG features, classified using an SVM. In [1], authors use SIFT features, compressed with LDA, and classify them using several machine learning techniques. Texture descriptors, such as oriented texture curves [19] and semantic proto-concepts [10] have also been adopted for general scene classification. To include spatial information in the holistic descriptors, several researchers apply spatial pyramid histograms [17, 14].

The second line of work employs object detectors and reasons about the room category based on occurrence of their responses. In [9], authors propose to process an input image with a large bank of pre-trained object detectors, and combine their responses into a single scene descriptor. The same filter bank is used by [11], who construct a high-level image representation based on co-occurrences of detector responses. Such approaches largely depend on the quality of used detectors, as well as their representativity for the particular room category. A more general approach based on local description was presented in [13]. It exploits both local and global visual information by describing the scene using multiple regions in a star-shaped model. Authors in [12] jointly learn the set of prototype regions and image classifiers by random sampling image regions; the final holistic representation is formed via concatenation of the obtained region responses. [15] train individual object detectors and object-group detectors, and max-pool their responses in a spatial pyramid. In a similar scheme, [5] discover visually-coherent patch clusters from an image collection that are maximally discriminative with respect to the labels.

The turning point in room classification has been the advent of large domain-specific evaluation datasets, which have provided a platform for method comparison and further development. Datasets such as MIT Indoor [13], SUN [20], Places205 [23], and Places365 [22] are among the most widely used. Some of them extend beyond indoor room categories. However, due to the holistic ground-truth labels, they are suitable only for evaluation of holistic room classification methods.

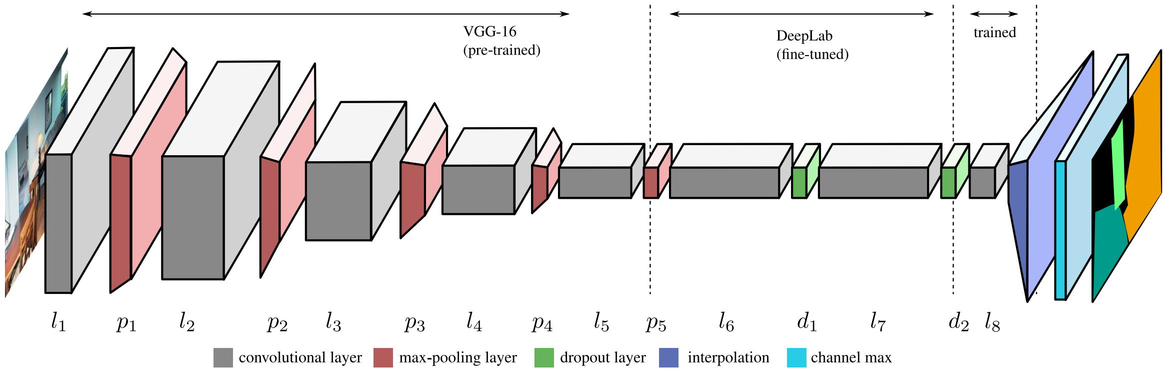


Figure 2: Schematic representation of the RoomNet network architecture.

The availability of large training datasets has also boosted the utilization of deep-learning techniques. In particular, convolutional neural networks have been successfully applied to scene recognition in [23]. The authors train the network on combination of ImageNet [4] and Places205 [23] datasets, and use it as a generic feature extractor in combination with an SVM to achieve state-of-the-art results on several scene recognition datasets. In [6], authors use cutouts at multiple scales as the input to networks in order to better handle geometric invariance. The system proposed in [21] generates activation maps that correspond to regions with discriminative information for each class, and uses them to classify images. Another part-based system [18] uses CNN features extracted from generic region proposals to increase robustness to partial occlusions. In contrast to our work, all these methods still address the holistic task of predicting a single label for the input image.

The use of segmentation in our work is reminiscent of the scene-parsing task [24]. The latter is low-level, and is concerned with segmenting the image into regions that correspond to individual objects. Therefore, our work can be seen as situated between the high-level holistic scene classification and low-level scene parsing, and instead of providing a single category label or segmented individual objects, aims to segment image regions that correspond to distinct semantic room types.

3. Methodology

The theoretical framework of our approach is the DeepLab [2] CNN architecture which was developed for generic semantic segmentation. The DeepLab network is based on the VGG-16 network [16], and is fine-tuned for segmentation of the 21 categories from Pascal VOC 2012 dataset. The bottom ten layers are based on VGG-16, comprising five series of convolutional layers with interspersed max-pooling layers. At the top of the network, there are three fully-connected layers, separated by two drop-out layers. The output of the final layer is a per-pixel label distribution map, which is spatially interpolated with factor 8 to the original image resolution. The DeepLab [2] originally applies a conditional random field (CRF) to improve per-pixel segmentation accuracy.

Our network, which we call the *RoomNet*, is depicted in Figure 2. The *RoomNet* re-uses the bottom ten convolutional and max-pool layers of the DeepLab network, and re-trains the top three fully-connected layers on the task of segmentation for room localization and categorization. The top fully-connected (output) layer, l_8, contains nine channels corresponding to the eight room categories in our dataset plus the *background* label. The ninth channel corresponding to the background class is used only in the training phase to account for the "unassigned" pixels in the ground truth. The network architecture and parameters of individual layers are summarized in Table 1.

3.1. Training

In the training stage, each spatial unit (collection of the channel responses at the same spatial position) is connected to a softmax classifier that predicts the most likely class label at that location. Due to reductions in the lower network layers, the resulting output map is subsampled by factor 8 compared to the input image size. The network upper layers are optimized with respect to the loss function, based on the sum of cross-entropy terms for each spatial unit, with all positions and labels being weighted equally, and the targets being the pixel-wise ground-truth labels (also

Table 1: Properties of the RoomNet network layers. The l_i layers are convolutional, p_i layers are max-pooling, and d_i layers are drop-out layers. In the upper part of the network, the fully-connected layers (l_6, l_7, and l_8) are implemented via convolutional layers. The last column contains filter parameters: kernel size K, padding P, and stride S. If not specified, $P = 0$ and $S = 1$. H denotes the parameter of the "atrous algorithm" introduced by [2].

Layer	Size	Channels	Filter parameters
l_1	256×256	64	$K = 3, P = 1$
p_1	129×129	64	$K = 2, P = 1, S = 2$
l_2	129×129	128	$K = 3, P = 1$
p_2	65×65	128	$K = 2, P = 1, S = 2$
l_3	65×65	256	$K = 3, P = 1$
p_3	33×33	256	$K = 2, P = 1, S = 2$
l_4	33×33	512	$K = 3, P = 1$
p_4	32×32	512	$K = 2, S = 1$
l_5	32×32	512	$K = 3, P = 2, H = 2$
p_5	32×32	512	$K = 3, P = 1, S = 1$
l_6	32×32	4096	$K = 4, P = 6, H = 4$
d_1	32×32	4096	
l_7	32×32	4096	$K = 1$
d_2	32×32	4096	
l_8	32×32	9	$K = 1$

subsampled by the factor 8). For details of the cost function, we refer the interested reader to [2].

3.2. Testing

During the inference, the response maps of the l_8 layer are interpolated to the original image size. After the interpolation, the final segmentation map is obtained by selecting a label with highest probability at each pixel location. In contrast to the DeepLab [2], we do not apply the CRF, since high per-pixel accuracy is not required for room category inference. The boundaries between different room types in images are not as well defined as the boundaries between objects, therefore each room category can be localized only approximately. Any further spatial regularization would merely result in an unnecessary increase in computational cost.

4. Experimental evaluation

The implementation details of our approach are provided in Section 4.1. The dataset is described in Section 4.2. The experimental results are described and discussed in Section 4.3.

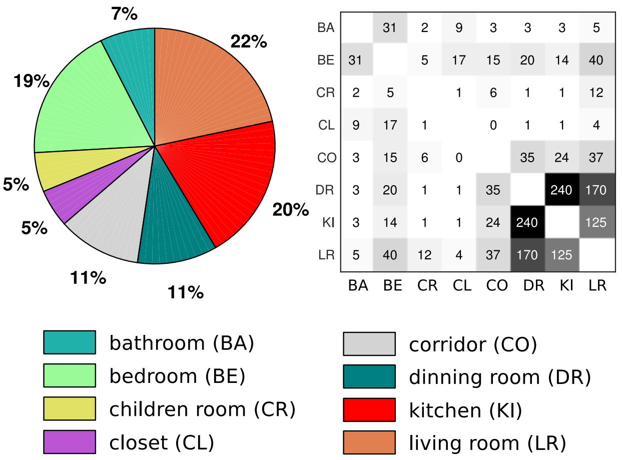
4.1. Implementation details

The *RoomNet* network was implemented in the Caffe framework [7]. We have used the training portion of the dataset, presented in Section 4.2, to train the top three fully-connected layers. The training hyperparameters were set to the values recommended by the Caffe documentation. Layers l_6 and l_7 are initialized with the DeepLab data and only fine-tuned for our scenario with learning rate 10^{-3}. Layer l_8 is initialized with random weights, and trained from scratch with initial learning rate 10^{-2}, which is afterwards decreasing with factor $\lambda = 0.1$ per epoch. We use mini-batch size of 10 images, training momentum $\mu = 0.9$, and weight decay of 0.0005.

4.2. Dataset

The existing datasets are tailored for evaluation of holistic room categorization performance and contain only a single category label per image. We thus assembled a new dataset that provides pixel-wise room category annotations. Our dataset is based on the MIT Indoor [13] dataset, which contains images of 67 indoor scene categories. Note that we have noticed that in fact several images from this dataset contain multiple room categories, but are annotated only with a single category. We selected a subset of 8 categories corresponding to most common household rooms: *bathroom*, *bedroom*, *children room*, *closet*, *corridor*, *dinning room*, *kitchen*, and *living room*. This selection was extended by additional images from the internet, that contained multiple room categories within a single image. To account for localization ambiguity, we introduced a special category for the *background*, which accounts for the image regions that cannot be reliably attributed to any of the room categories.

The dataset is summarized in Figure 3. It contains 8029 images, 1778 of which are in the testing set. We have made sure that the distribution of classes in both training and testing sets are equal. In total 465 images contain two or more room categories. Each image is manually annotated on per-pixel basis with the nine possible category labels. From the co-occurrence matrix in Figure 3 we can see that most categories appear together at least in one image. The most frequently co-occurring rooms are kitchen, dining room, and living room. Such arrangements are typical for small apartments.



	BA	BE	CR	CL	CO	DR	KI	LR
BA		31	2	9	3	3	3	5
BE	31		5	17	15	20	14	40
CR	2	5		1	6	1	1	12
CL	9	17	1		0	1	1	4
CO	3	15	6	0		35	24	37
DR	3	20	1	1	35		240	170
KI	3	14	1	1	24	240		125
LR	5	40	12	4	37	170	125	

Figure 3: Our dataset overview. The pie-chart on the left presents the distribution of room categories, the co-occurrence matrix on the right shows the number of simultaneous appearances of two categories in the same image.

4.3. Experimental results

We compare the proposed approach against the state-of-the-art room categorization method called the *PlacesNet* from [23]. The PlacesNet applies a pre-trained CNN as a generic feature extractor to encode the input image as a 4096-dimensional feature vector and an SVM for classification. We consider the variant with Places205-AlexNet network, which was trained on the Places205 [23] dataset. To train this pipeline on our dataset, we extract the CNN features from training images with the pre-trained Places205-AlexNet, and use them to train the linear one-vs-rest SVM. Note that we train only with the subset of training images that contain a single room category. In our preliminary study we trained with multi-room samples as well (i.e., treating each such image as multiple samples with corresponding labels), but have observed significantly worse performance.

Two evaluation scenarios are considered. The first scenario considers room category presence detection. Here, the task is to correctly identify the labels for the categories present in the image. The second scenario considers localization of these categories as well. In addition we conduct a robustness study in which the input images are rotated. Image rotation is common in mobile platforms due to camera tilting during robot motion.

Table 2: Summary of the experimental results. The table shows the accuracy for single-room classification task, the best F-score for multi-room classification task, and the IoU score for the localization task.

Approach	Accuracy	F-score	IoU
PlacesNet [23]	0.79	0.72	0.73
RoomNet (ours)	**0.90**	**0.83**	**0.84**

4.4. Room category presence detection

To maintain relation to the existing holistic benchmarks, both methods were first compared in terms of classification accuracy on a subset of images that only contain a single room category.

The most likely category is determined in our RoomNet by determining the category with maximum number of pixels associated. The accuracy of classification is shown in the left-most column of Table 2. Our RooomNet significantly outperforms the PlacesNet even on this, traditionally holistic, task by approximately $14\,\%$. This boost speaks in favor of our non-parametric formulation of category estimation through segmentation. The ambiguous pixels, or pixels stemming potentially from other categories, represent non-dominant categories and do not affect dominant class prediction, thus increasing the robustness of RoomNet.

Next, the entire dataset was considered for multi-category classification. Both methods were adapted to report classes where the reported score of the class is higher than the Θ portion of the maximum output class. For PlacesNet the class score is the output of each one-vs-all SVM classifier (distance to the decision hyperplane) normalized by softmax to obtain the presence probability of each class. For RoomNet the probability of each class corresponds to the percentage of the pixels assigned to it.

The multi-category results are shown in Figure 4 in terms of F-score and precision-recall with respect to the threshold Θ. The RoomNet consistently outperforms the PlacesNet in F-score. In fact, the RoomNet achieves an excellent performance already at moderate threshold levels. Since our dataset is dominated by single-category images, we show the F-scores separately for single-category and multiple-category images (second row in Figure 4). The single-category graph is similar to the overall F-score plot. The multiple-category graphs shows that our RoomNet outperforms the PlacesNet when a single category is reported ($\Theta = 1$). The improvement is most significant for Θ selected for each method separately. Table 2 depicts average performance of RoomNet and PlacesNet at their best values of Θ on single and mixed room category images. The RoomNet outperforms the PlacesNet by over $15\,\%$. The precision-recall curve of RoomNet in Figure 4 is consistently above that of the PlacesNet. The best recall achieved by PlacesNet is close to 0.8, with 0.68 precision. Our RoomNet achieves approximately 0.84 precision at the 0.8 recall, thus by far outperforming the PlacesNet.

The results again support our non-parametric formulation in which the category of a single pixel does not pollute the global classification score, but rather contributes to the specific class through a cluster of pixels.

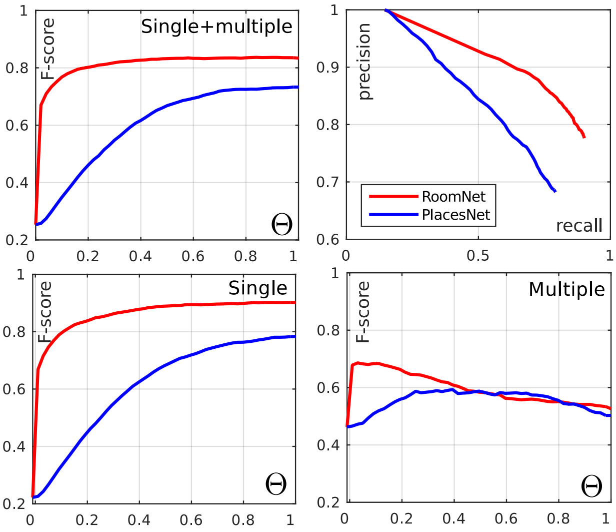


Figure 4: Top row: F-score and precision-recall curves for multi-class categorization with respect to parameter Θ. Bottom row: F-score plots with respect to parameter Θ only for single room and multiple room images.

4.5. Room category localization

The task of room category localization in an image is posed as a semantic segmentation problem. Similarly to semantic segmentation evaluation, we estimate the localization performance by the intersection-over-union (IoU) score. The score is modified to account for the background class that we have reserved for ambiguous regions in the ground truth annotation. Pixels that are labeled as background in the ground truth are ignored in the calculation of the IoU score since the correct class is unknown at these locations.

Holistic methods are unable to provide category location information, since they provide only a single output label. Thus, the segmentation map for PlacesNet is computed by assigning the output label to all pixels in the image. Such straightforward extension indeed results in suboptimal performance on multi-room images as only a single room category is reported. On the other hand, PlacesNet should have advantage on images that contain only a single class, since a correct global decision results in $100\,\%$ labeling accuracy. The RoomNet is required to make such predictions for each pixel independently, which is a competitive disadvantage in single-category images. This is illustrated in Figure 5. The PlacesNet method predicts more images with a perfect IoU score, but those are all single-category images. It also receives a very low score on a large portion of the testing dataset. These are the images with multiple rooms and mis-categorized single rooms where the IoU is $0\,\%$. The RoomNet method achieves high (although sometimes not perfect) IoU in single-category images and achieves a better global performance than the PlacesNet. According to the average IoU score in Table 2, the RoomNet outperforms the PlacesNet by $15\,\%$ in IoU score.

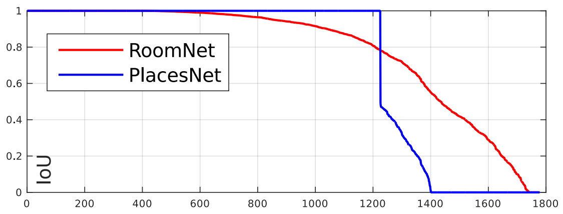


Figure 5: IoU score for all testing images sorted in decreasing order.

Figure 6 shows confusion matrices computed from pixel-wise classifications. Our RoomNet outperforms PlacesNet across all categories. The PlacesNet sometimes miss-classifies bathroom and childrens room for bedroom. The RoomNet significantly reduces these mixups. The reduction in false classifications is also apparent in the case of dinning rooms. PlacesNet often miss-classifies them as a kitchen or a living room. The miss-classification is reduced by RoomNet by 15 percentage points ($\sim 20\,\%$ improvement).

4.6. Robustness to image rotation

To evaluate the rotational invariance we have repeated the classification and localization experiments

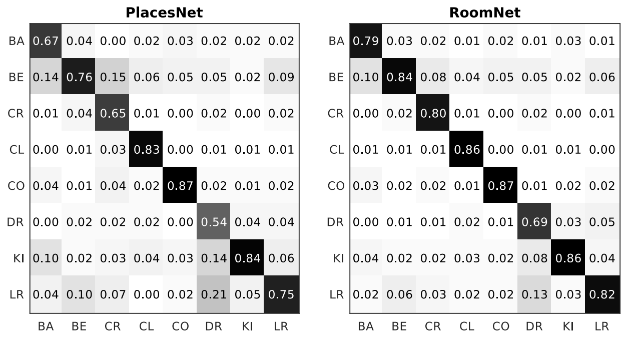


Figure 6: Confusion matrices for localization task.

from Section 4.4 and Section 4.5 on a modified datasets where all images were rotated by an angle $\alpha \in [0, 180]$ degrees.

The results for the single and multiple room classification are shown in Figure 8. Results show a remarkable robustness of RoomNet to rotations of up to 30 degrees. The drop in accuracy is within 10 %, and approximately 5 % in F-score and IoU. In case of PlacesNet, the drop in accuracy is twice as large, while in F-score and IoU is four times as large. At 30 degree rotation, the PlacesNet accuracy is approximately 0.59, while the accuracy of our RoomNet is still at approximately 0.83. Thus the improvement of our RoomNet over PlacesNet is in order of 40 % at 30 degree rotation. For significant rotations, the peformance of both approaches drops, but the drop is still more significant for the PlacesNet than for our RoomNet.

The segmentation performance for different image rotation angles is visualized in Figure 8. On single room images the holistic PlacesNet method has the advantage as it estimates only global category, however, if the method fails, the entire segmentation is incorrect. The RoomNet produces some small incorrect regions in such cases because it has to classify them individually, but overall the results are consistent with the ground truth labels. In case of multi-room images the difference is even more apparent. PlacesNet can only recognize a single category, usually the dominant one. But in many cases the results are completely incorrect. The RoomNet provides a decent localization. For unusual rotation angles, the segmentation also includes small regions of incorrect predictions that could be improved in a post-processing step which takes into account the size of the regions and prior information about co-occurrence of room categories.

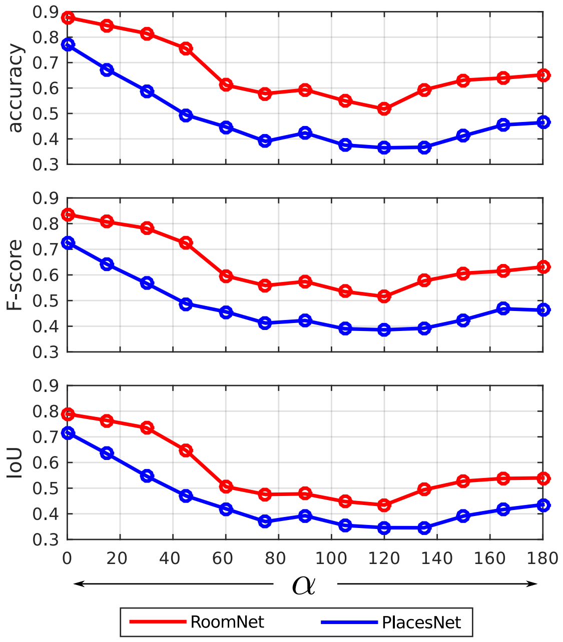


Figure 7: Classification accuracy for single room classification task and F-score for multi-room classification in relation to input image rotation.

5. Conclusion

A new approach for image-based room categorization was proposed. In contrast to related approaches that perform a global classification by aggregating features into a holistic representation, our approach applies a non-parametric technique. The non-parametric property comes from estimating the support for a particular category at a pixel level and separately aggregating the votes for each category. This is achieved by casting the categorization problem as a semantic segmentation task in which each pixel is assigned a room category label. A CNN framework is used to construct our semantic segmentation network – the RoomNet. In addition to the novel room recognition network we presented a new dataset for room categorization and localization. The dataset is per-pixel annotated groundtruth with eight room categories.

Extensive analysis was performed to compare the RoomNet with a recent state-of-the-art PlacesNet [23]. Our approach outperforms the PlacesNet on the task of single-category as well as multi-category classification. The RoomNet exhibits an excellent ability for localization of categories in the images and is significantly more robust to image rotation than the PlacesNet. The RoomNet achieves approximately 15 % better accuracy in room category

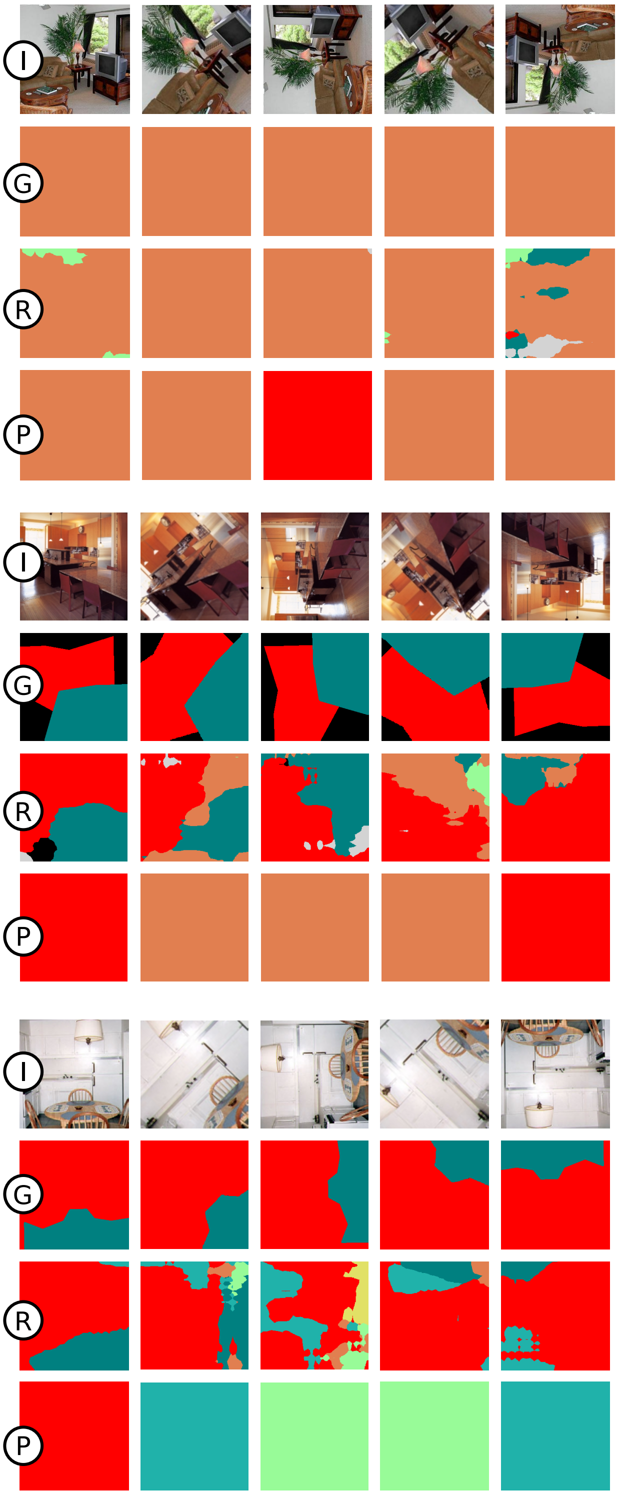


Figure 8: Examples of segmentation obtained for different rotation angles. The top example shows performance for a single room image, while the other two examples show performance for images where two rooms are present in the same image.

detection. Our RoomNet goes beyond the state-of-the-art by allowing room category localization in addition to recognition.

Our future work will include extending the RoomNet by increasing the (indoor and outdoor) place categories. The approach will be extended to a mobile robot scenario in which the provided segmentations from RoomNet will be used to increase the accuracy of localization and motion planning during robotic exploration.

References

[1] B. Ayers and M. Boutell. Home interior classification using SIFT keypoint histograms. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–6. IEEE, 2007. 2

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Semantic image segmentation with deep convolutional nets and fully connected CRFs. *arXiv preprint arXiv:1412.7062*, 2014. 3, 4

[3] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004. 2

[4] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*, pages 248–255, 2009. 3

[5] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 494–502, 2013. 2

[6] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *European Conference on Computer Vision*, pages 392–407. Springer, 2014. 1, 3

[7] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM, 2014. 4

[8] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 923–930, 2013. 2

[9] L.-J. Li, H. Su, L. Fei-Fei, and E. P. Xing. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In *Advances in neural information processing systems*, pages 1378–1386, 2010. 1, 2

[10] R. Margolin, L. Zelnik-Manor, and A. Tal. OTC: A novel local descriptor for scene classification. In *European Conference on Computer Vision*, pages 377–391. Springer, 2014. 2

[11] G. Mesnil, S. Rifai, A. Bordes, X. Glorot, Y. Bengio, and P. Vincent. Unsupervised learning of semantics of object detections for scene categorization. In *International Conference on Pattern Recognition Applications and Methods (ICPRAM 2015)*, pages 209–224, 2015. 2

[12] S. N. Parizi, A. Vedaldi, A. Zisserman, and P. Felzenszwalb. Automatic discovery and optimization of parts for image classification. *arXiv preprint arXiv:1412.6598*, 2014. 2

[13] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 413–420. IEEE, 2009. 1, 2, 4

[14] F. Sadeghi and M. F. Tappen. Latent pyramidal regions for recognizing scenes. In *European Conference on Computer Vision*, pages 228–241. Springer, 2012. 2

[15] A. Sadovnik and T. Chen. Hierarchical object groups for scene classification. In *19th IEEE International Conference on Image Processing (ICIP 2012)*, pages 1881–1884, 2012. 2

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 3

[17] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *European Conference on Computer Vision*, pages 73–86. Springer, 2012. 2

[18] P. Uršič, R. Mandeljc, A. Leonardis, and M. Kristan. Part-Based Room Categorization for Household Service Robots, 2016. 3

[19] J. C. van Gemert, J. Geusebroek, C. J. Veenman, C. G. M. Snoek, and A. W. M. Smeulders. Robust scene categorization by learning image statistics in context. In *2006 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06)*, page 105. IEEE, 2006. 2

[20] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Computer vision and pattern recognition (CVPR), 2010 IEEE conference on*, pages 3485–3492. IEEE, 2010. 2

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning Deep Features for Discriminative Localization. *arXiv preprint arXiv:1512.04150*, 2015. 1, 3

[22] B. Zhou, A. Khosla, A. Lapedriza, A. Torralba, and A. Oliva. Places: An image database for deep scene understanding. *arXiv preprint arXiv:1610.02055*, 2016. 2

[23] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in neural information processing systems*, pages 487–495, 2014. 1, 2, 3, 5, 7

[24] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba. Semantic understanding of scenes through the ade20k dataset. *arXiv preprint arXiv:1608.05442*, 2016. 3