



# Multi-stream Convolutional Networks for Indoor Scene Recognition

Rao Muhammad Anwer<sup>1,2(✉)</sup>, Fahad Shahbaz Khan<sup>2,3(✉)</sup>,  
Jorma Laaksonen<sup>1(✉)</sup>, and Nazar Zaki<sup>4(✉)</sup>

<sup>1</sup> Department of Computer Science, Aalto University School of Science,  
Espoo, Finland

`{rao.anwer,jorma.laaksonen}@aalto.fi`

<sup>2</sup> Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

<sup>3</sup> Computer Vision Laboratory, Linköping University, Linköping, Sweden  
`fahad.khan@liu.se`

<sup>4</sup> Computer Science and Software Engineering Department,  
College of Information Technology, United Arab Emirates University, Al Ain, UAE  
`Nzaki@uaeu.ac.ae`

**Abstract.** Convolutional neural networks (CNNs) have recently achieved outstanding results for various vision tasks, including indoor scene understanding. The de facto practice employed by state-of-the-art indoor scene recognition approaches is to use RGB pixel values as input to CNN models that are trained on large amounts of labeled data (ImageNet or Places). Here, we investigate CNN architectures by augmenting RGB images with estimated depth and texture information, as multiple streams, for monocular indoor scene recognition. First, we exploit the recent advancements in the field of depth estimation from monocular images and use the estimated depth information to train a CNN model for learning deep depth features. Second, we train a CNN model to exploit the successful Local Binary Patterns (LBP) by using mapped coded images with explicit LBP encoding to capture texture information available in indoor scenes. We further investigate different fusion strategies to combine the learned deep depth and texture streams with the traditional RGB stream. Comprehensive experiments are performed on three indoor scene classification benchmarks: MIT-67, OCIS and SUN-397. The proposed multi-stream network significantly outperforms the standard RGB network by achieving an absolute gain of 9.3%, 4.7%, 7.3% on the MIT-67, OCIS and SUN-397 datasets respectively.

**Keywords:** Scene recognition · Depth features · Texture features

## 1 Introduction

Scene recognition is a fundamental problem in computer vision with numerous real-world applications. The problem can be divided into recognizing indoor

versus outdoor scene types. Initially, most approaches target the problem of outdoor scene classification with methods demonstrating impressive performance on standard benchmarks, such as fifteen scene categories [17]. Later, the problem of recognizing indoor scene categories have received much attention with the introduction of specialized indoor scene datasets, including MIT-67 [23]. Different to outdoor scene categorization, where global spatial layout is distinctive and one of the most discriminative cues, indoor scenes are better characterized either based on global spatial properties or local appearance information depending on the objects they contain. In this work, we investigate the challenging problem of automatically recognizing indoor scene categories.

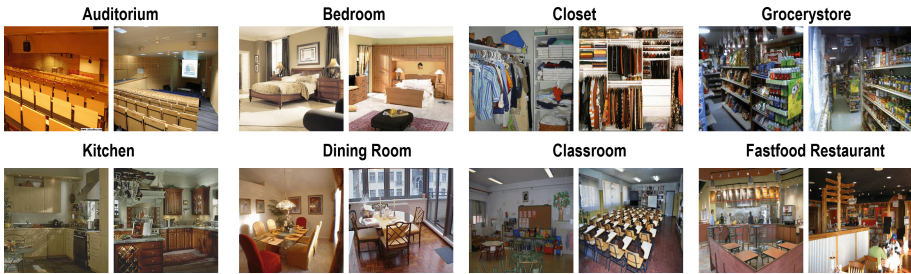
In recent years, deep convolutional neural networks (CNNs) have revolutionized the field of computer vision setting new state-of-the-art results in many applications, including scene recognition [32]. In the typical scenario, deep networks or CNNs take raw pixel values as an input. They are trained using a large amount of labeled data and perform a series of convolution, local normalization and pooling operations (called layers). Generally, the final layers of a deep network are fully connected (FC) and employed for the classification purpose. Initially, deep learning based scene recognition approaches employed CNNs pre-trained on the ImageNet [26] for object recognition task. These pre-trained deep networks were then transferred for the scene recognition problem. However, recent approaches have shown superior results when training deep networks on a specialized large-scale scene recognition dataset [32]. In all cases, the de facto practice is to use RGB patches as input when training these networks.

As mentioned above, the standard procedure is to employ RGB pixel values as input for training deep networks. Besides color, texture features also provide a strong cue for scene identification at both the superordinate and basic category levels [24]. Significant research efforts have been dedicated in the past in designing discriminative texture features. One of the most successful hand-crafted texture descriptors is that of Local Binary Patterns (LBP) and its variants [12, 21, 22]. LBP is based on the signs of differences of neighboring pixels in an image and is invariant to monotonic gray scale variations. Recent studies [1, 4] have investigated employing deep learning to design deep texture representations.

Other than color and texture, previous works [7, 11, 27, 28] have shown the effectiveness of depth information and that depth images can be used simultaneously with RGB images to obtain improved recognition performance. However, most of these approaches require depth data acquired from depth sensors together with camera parameters to associate point clouds to image pixels. Despite increased availability of RGB-D sensors, standard large-scale object and scene recognition benchmarks (ImageNet and Places) still contain RGB images captured using different image sensors with no camera parameters to generate accurate point clouds. In a separate research line, recent works [5, 20] have investigated estimating depth information from single monocular images. These methods employ RGB-D acquired through depth sensors during the training stage to infer the depth of each pixel in a single RGB image. Here, we aim

to exploit these advancements in depth estimation from monocular images *and* hand-crafted discriminative texture features to integrate explicit depth and texture information for indoor scene recognition in the deep learning architecture.

In this work, we propose a multi-stream deep architecture where the estimated depth and texture streams are fused with the standard RGB image stream for monocular indoor scene recognition. The three streams can be integrated at different stages in the deep learning architecture to make use of the complementary information available in these different modalities. In the first strategy, the three streams are integrated at an early stage by aggregating the RGB, texture and estimated depth image channels as the input to train a joint multi-stream deep CNN model. In the second strategy, the three streams are trained separately and combined at a later stage of the deep network. To the best of our knowledge, we are the first to propose a multi-stream deep architecture and investigate fusion strategies to combine RGB, estimated depth and texture information for monocular indoor scene recognition. Figure 1 shows example indoor scene categories from the MIT-67 dataset and their respective classification accuracies (in %) when using different streams and their combination in the proposed multi-stream architecture.



|              | Auditorium | Bedroom   | Closet    | Grocerystore | kitchen   | Dining Room | Classroom | Fastfood Restaurant |
|--------------|------------|-----------|-----------|--------------|-----------|-------------|-----------|---------------------|
| RGB          | 33         | 57        | 72        | 62           | 52        | 44          | 67        | 53                  |
| Depth        | 56         | 48        | 83        | 48           | 56        | 33          | 61        | 35                  |
| Texture      | 56         | 61        | 72        | 52           | 57        | 50          | 72        | 35                  |
| Three-Stream | <b>72</b>  | <b>67</b> | <b>84</b> | <b>86</b>    | <b>71</b> | <b>78</b>   | <b>83</b> | <b>94</b>           |

**Fig. 1.** Example categories from MIT-67 indoor scene dataset and their respective classification accuracies (in %) when using different streams: baseline standard RGB, estimated depth and texture. We also show the classification accuracies when combining these streams in our late fusion based three-stream architecture. The classification results are consistently improved with our three-stream architecture, highlighting the complementary information possessed by the three streams.

## 2 Related Work

**Indoor Scene Recognition:** Recently, indoor scene recognition has gained a lot of attention [6, 8, 14–16]. Koskela [16] propose an approach where CNNs, trained on object recognition data, using different architectures are employed as

feature extractors in a standard linear-SVM-based multi-feature scene recognition framework. A discriminative image representation based on discriminative mid-level convolutional activations is proposed by [14] to counter variability in indoor scenes. Guo et al. [6] propose an approach by integrating local convolutional supervision layer that is constructed upon the convolutional layers of deep network. The work of [15] proposes an approach based on spectral transformation of CNN activations integrated as a unitary transformation within a deep network. All these aforementioned deep learning based approaches are trained using RGB pixel values of an image.

**Depth Estimation:** Recent approaches [5, 19, 20] employ deep learning to learn depth estimation in monocular images. The work of [5] proposes a multi-scale convolutional architecture for depth prediction, surface normals and semantic labeling. Li et al. [19] introduce an approach by regressing CNN features together with a post-processing refinement step employing conditional random fields (CRF) for depth estimation. The work of [20] proposes a deep convolutional neural field model that jointly learns the unary term and pairwise term of continuous CRF in a unified CNN framework. Different to [5], where the depth map is directly regressed via convolutions from an input image, the approach of [20] explicitly models the relations of neighbouring superpixels by employing CRF. Both unary and binary potentials are learned in a unified deep network framework. Here, we employ deep convolutional neural field model of [20] as a depth estimation strategy for our monocular deep depth network stream. In our multi-stream architecture, the monocular depth stream is trained from scratch, on the large-scale ImageNet and Places datasets, for indoor scene recognition.

**Texture Representation:** Robust texture description is one of the fundamental problems in computer vision and is extensively studied in literature. Among existing methods, the Local Binary Patterns (LBP) descriptor [22] is one of the most popular hand-crafted texture description methods and several of its variants have been proposed in literature [21]. Recent approaches [1, 4] have investigated deep learning for the problem of texture description. Cimpoi et al. [4] propose to encode convolutional layers of the deep network using the Fisher Vector scheme. Rao et al. [1] investigate the problem of learning texture representation and integrate LBP within deep learning architecture. In that approach, LBP codes are mapped to points in a 3D metric space using the approach of [18]. Here, we employ the strategy proposed in [1] to learn the texture stream and combine it with RGB and estimated depth streams in a multi-stream deep architecture for indoor scene recognition.

### 3 Our Multi-stream Deep Architecture

Here, we present our multi-stream deep architecture for indoor scene recognition. We also investigate fusion schemes to integrate different modalities in the deep learning architecture. We base our approach on the VGG architecture [3] that takes as input an image of  $224 \times 224$  pixels and consists of five convolutional (conv) and three fully-connected (FC) layers.

### 3.1 Deep Depth Stream

The first step in designing of the depth stream is to compute the estimated depth image given its RGB counterpart. We employ the method of [20] for depth estimation of each pixel in a monocular image. The depth estimation approach employs continuous CRF to explicitly model the relations of neighbouring superpixels. Both unary and binary terms of continuous CRF are learned in an unified deep network framework. In the depth estimation model, each image is comprised of small regions, termed as superpixels, with nodes of a graphical model defined on them. Each superpixel in an image is described by the depth value of its centroid. Let  $I$  be an image and  $y = [sp_1, \dots, sp_m]^\top \in \mathbb{R}^m$  be a vector of all  $m$  superpixels in image  $I$ . The conditional probability distribution of the data is then modelled by employing the following density function:

$$P(y | I) = \frac{1}{Z(I)} \exp(-EN(y, I)), \quad (1)$$

where  $EN$  is the energy function and the partition function represented by  $Z$  is defined as:

$$Z(I) = \int_y \exp\{-EN(y, I)\} dy. \quad (2)$$

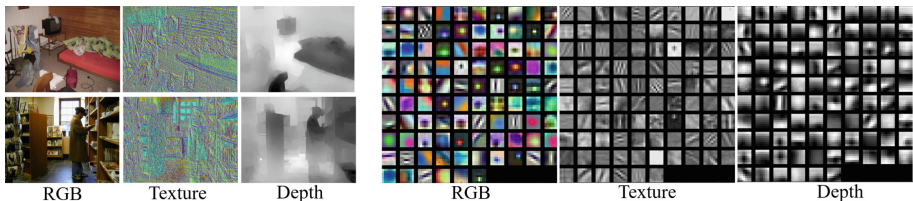
Due to the continuous nature of the depth values  $y$ , no approximation method is required to be applied. The subsequent MAP inference problem is then solved in order to obtain the depth value of a new image. The energy function is written as a combination of unary potentials  $UN$  and pairwise potentials  $PV$  over the superpixels  $\mathcal{M}$  and edges  $\mathcal{S}$  of the image  $I$ :

$$EN(y, I) = \sum_{p \in \mathcal{M}} UN(y_p, I) + \sum_{(p, q) \in \mathcal{S}} PV(y_p, y_q, I), \quad (3)$$

Here, the unary potential  $UN$  regresses the depth value for a single superpixel whereas the pairwise potential  $PV$  invigorates the superpixel neighborhoods with similar appearances to hold similar depth values. In the work of [20], both the unary potentials  $UN$  and the pairwise potentials  $PV$  are learned jointly in a unified deep network framework. The deep network comprises the following components: a continuous CRF loss layer consisting of a unary part and a pairwise part. Given an input image, image patches centred around each superpixel centroid are considered. Each image patch is used as an input to the unary part which is fed into the deep network. The network returns a single value representing the regressed depth value of the superpixel. The unary part of the deep network consists of five convolutional and four fully-connected layers. The unary potential is formulated by the output of the deep network by considering the following least square loss:

$$UN(y_p, I; \theta) = (y_p - z_p(\theta))^2, \forall p = 1, \dots, m, \quad (4)$$

Here,  $z_p$  is the regressed depth of the homogeneous region (superpixel)  $p$ , parameterized by the deep network parameters  $\theta$ . In case of the pairwise part of the network, the input is the similarity vectors of all neighboring superpixel pairs, fed to the FC layer with shared parameters among different superpixel pairs. The pairwise term enables neighboring superpixels with similar appearances to have similar depth values. Three types of pairwise similarities are considered: color histogram difference, color difference and texture disparity based on LBP. The output is then a 1-dimensional similarity vector for each of the neighboring superpixel pairs. Consequently, outputs from the unary and the pairwise terms are taken by the continuous CRF loss layer in order to minimize the negative log-likelihood. Standard RGB-D datasets, including NYUD2 have the same viewing angles for both the camera and the depth sensor. This implies that objects in a depth image possess the same 2D shapes as in RGB image with the only difference is that the RGB values are replaced by depth values. The estimated depth images alleviate the problem of intra-object variations, which is desired for scene understanding. During the construction of the depth stream, we first estimate depth values of the input RGB image using the approach described above resulting in a single-channel depth map. The estimated depth values are log-normalized to the range of  $[0, 255]$  and duplicated into three channels which are then input to the deep learning framework. Figure 2 shows example RGB images and their corresponding estimated depth maps.

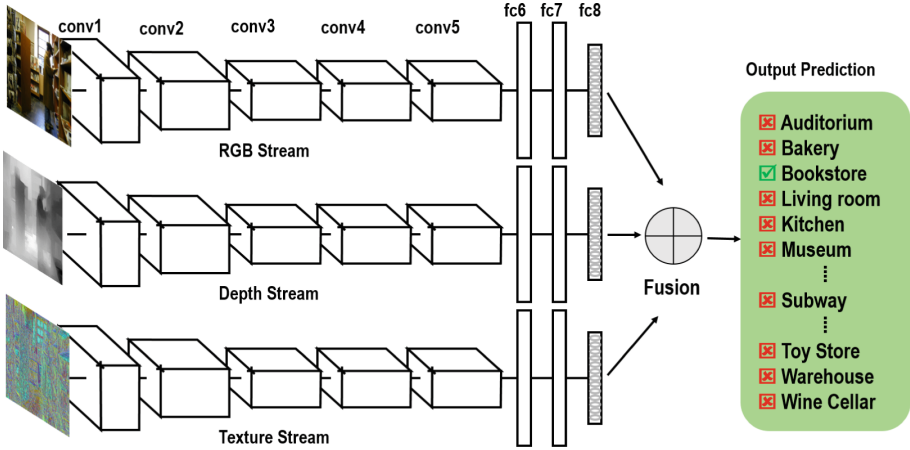


**Fig. 2.** On the left: example RGB images and the corresponding texture coded mapped images (visualized here in color) together with estimated depth images. On the right: visualization of filter weights from the RGB, texture and estimated depth CNN models.

### 3.2 Deep Texture Stream

In addition to the standard RGB and estimated depth streams, we propose to integrate an explicit texture stream for indoor scene recognition since texture features have shown to be crucial for scene understanding. Here, we base our texture stream on the popular LBP descriptor [22] where the neighborhood of a pixel is described by its binary derivatives used to form a short code for the neighborhood description of the pixel. These short codes are binary numbers (lower than threshold (0) or higher than the threshold (1)), where each LBP code can be regarded as a micro-texton. Each pixel in the image is allocated a code of the texture primitive with its best local neighborhood match.

When integrating the LBP operator in the deep learning architecture, a straightforward way is to directly employ LBP codes as an input to the deep network. However, the direct incorporation of LBP codes as input is infeasible since the convolution operations, equivalent to a weighted average of the input values, employed within CNNs are unsuitable for the unordered nature of the values of the LBP code. To counter this issue, the work of [18] proposes to map the LBP code values to points in a 3D metric space. In this metric space, the Euclidean distance approximates the distance between the LBP code values. Such a transformation enables averaging of LBP code values during convolution operations within CNN models. First, a distance  $\delta_{j,k}$  is defined between the LBP codes  $LBPT_j$  and  $LBPT_k$ . In the work of [18], Earth Mover's Distance (EMD) [25] is employed since it takes into account both the different bit values and their locations. Afterwards, a mapping is derived of the LBP codes into a  $DM$ -dimensional space which approximately preserves the distance between them. The mapping is derived by applying Multi Dimensional Scaling (MDS) [2]. The mapping enables the transfer of LBP code values into a representation that is suitable to be used as input to the deep network. As in [1, 18], the dimensionality  $DM$  is set to three and the resulting texture representation is used to train a texture stream for indoor scene recognition. Figure 2 shows example RGB images and their corresponding texture coded mapped images.



**Fig. 3.** Our late fusion based multi-stream deep architecture. In this architecture, RGB, estimated depth and texture streams are kept separate and the point of fusion, which combines the three network towers, is at the end of the network.

### 3.3 Multi-stream Fusion Strategies

We consider two fusion strategies to integrate the RGB, estimated depth and texture streams in a multi-stream architecture. In the first strategy, termed as



early fusion, the three network streams are combined at an early stage as inputs to the deep network. As a result, the input to CNN is of  $224 \times 224 \times N$  dimensions, where  $N$  is the number of image channels. When combining the three streams in an early fusion strategy, the number of image channels is  $N = 7$  (3 RGB, 1 depth and 3 texture). A joint deep model is trained due to the aggregation of the image channels. In the second fusion strategy, termed as late fusion, the three networks are trained separately. The standard RGB stream network takes raw RGB values as input. The texture stream network takes texture coded mapped image as an input to the CNN model. This texture coded mapped image is obtained by first computing the LBP encoding that transforms intensity values in an image to one of the 256 LBP codes. The code values are then mapped into a 3D metric space, as described above, resulting in a 3-channel texture coded mapped image. The depth image is obtained by converting an RGB image to an estimated depth map, based on the procedure described earlier, to be used as an input to the depth stream. Consequently, the three streams are fused at the final stages of the deep network either by using FC layer activations with linear SVMs or combining the score predictions from individual streams. Figure 3 shows our late fusion based three stream architecture. The three streams are separately trained, from scratch, on both ImageNet [26] and Places [32] datasets. Figure 2 shows the VGG architecture based visualization of filter weights from the RGB, texture and estimated depth models trained on the ImageNet.

## 4 Experimental Results

**Experimental Setup:** We train our multi-stream network, described in Sect. 3, from scratch on the ImageNet 2012 [26] and Places 365 [32] training sets, respectively. In all cases, the learning rate is set to 0.001. The weight decay which contributes reducing the training error of the deep network is set to 0.0005. The momentum rate which is associated with the gradient descent method employed to minimize the objective function is set to 0.9. In case of fine-tuning the pre-trained deep models, we employ training samples with a batch size of 80, a momentum value of 0.9 and an initial learning rate of 0.005. Furthermore, in all experiments the recognition results are reported as the mean classification accuracy over all scene categories in a scene recognition dataset. From the network prediction, the scene category label providing the highest confidence is assigned to the test image. The overall results are obtained by calculating the mean recognition score over all scene classes in each scene recognition dataset.

**Datasets:** **MIT-67** [23] consists of 15,620 images of 67 indoor scene categories. We follow the standard protocol provided by the authors [23] by using 80 images per scene category for training and another 20 images for testing. **OCIS** [14] is the recently introduced large-scale object categories in indoor scenes dataset. It comprises of 15,324 images spanning more than 1300 commonly encountered indoor object categories. We follow the standard protocol provided by the authors [14] by defining a train-test split of (67% vs 33%) for each category. **SUN-397** [30] dataset consists of 108,754 images of 397 scene categories.



Here, the scene categories are both from indoor and outdoor environments. Each category in this dataset has at least 100 images. We follow the standard protocol provided by the authors [30] by dividing the dataset into 50 training and 50 test images per scene category. Since our aim is to investigate indoor scene recognition, we focus on the 177 indoor scene categories for the baseline comparison. Later, we show the results on the full SUN-397 dataset for state-of-the-art comparison.

**Baseline Comparison:** We compare our three-stream approach with the baseline standard RGB stream. Further, both early and late fusion strategies are evaluated for fusing the RGB, estimated depth and texture streams. For a fair comparison, we employ the same network architecture together with the same set of parameters for both the standard RGB and our multi-stream networks. Table 1 shows the baseline comparison with deep models trained on both ImageNet and Places datasets. We first discuss the results based on deep models pre-trained on the ImageNet. The baseline standard RGB deep network achieves average classification scores of 63.0%, 39.1%, and 46.0% on the MIT-67, OCIS, and SUN-397 datasets, respectively. The estimated depth based deep stream obtains mean recognition rates of 41.0%, 25.2%, and 26.0% on the MIT-67, OCIS and SUN-397 datasets, respectively. The texture coded deep image stream yields average classification accuracies of 59.1%, 33.6%, and 38.9% on the three scene datasets. In the case of the two fusion strategies, superior results are obtained with late fusion. The late fusion based two-stream network with RGB and depth streams obtains average classification scores of 67.1%, 40.9%, and 48.4% on the MIT-67, OCIS and SUN-397 datasets, respectively. Further, the late fusion based two-stream network with RGB and texture streams achieves average recognition rates of 69.3%, 42.5%, and 51.1% on the MIT-67, OCIS and SUN-397 datasets, respectively. The proposed late fusion based three-stream deep network significantly outperforms the baseline standard RGB deep stream on all datasets. Significant absolute gains of 9.3%, 4.7%, and 7.3% is achieved on the MIT-67, OCIS and SUN-397 datasets, respectively.

Other than the OCIS dataset, results are improved overall when employing deep models pre-trained on the Places scene dataset. The inferior recognition results in the case of the OCIS dataset are likely due to the fact that this dataset is based on indoor objects as categories instead of scenes. When comparing models trained on the Places dataset, our late fusion based three-stream deep architecture provides a substantial gains of 7.6%, 5.7%, and 4.9% on the MIT-67, OCIS and SUN-397 datasets respectively, compared to the baseline RGB stream.

We further analyze the impact of integrating depth and texture information within the deep learning architecture by looking into different indoor scene hierarchies available in the SUN-397 dataset. The indoor categories in the SUN-397 dataset are further annotated with the following scene hierarchies: shopping/dining with 40 indoor scene classes, workplace (office building, factory, lab, etc.) with 40 indoor scene classes, home/hotel with 35 indoor scene classes, transportation (vehicle interiors, stations, etc.) with 21 indoor scene classes, sports/leisure with 22 indoor scene classes, and cultural (art, education, religion,

**Table 1.** Comparison (overall accuracy in %) of our proposed three-stream deep architecture with the baseline standard RGB stream on the three scene datasets. We show multi-stream results with both early and late fusion schemes using deep networks either pre-trained on ImageNet or Places. Our proposed late-fusion based three-stream architecture significantly outperforms the baseline standard RGB stream on *all* datasets.

| Architecture                                      | Pre-training: imagenet |             |             | Pre-training: places |             |             |
|---|------------------------|-------------|-------------|----------------------|-------------|-------------|
|   | MIT-67                 | OCIS        | SUN-397     | MIT-67               | OCIS        | SUN-397     |
| RGB deep stream (baseline)                        | 63.0                   | 39.1        | 46.0        | 73.6                 | 32.5        | 58.6        |
| Depth deep stream                                 | 41.0                   | 25.2        | 26.0        | 51.5                 | 21.4        | 34.6        |
| Texture deep stream                               | 59.1                   | 33.6        | 38.9        | 68.7                 | 27.2        | 49.3        |
| Two-stream {RGB, depth} (early fusion)            | 65.2                   | 39.5        | 46.7        | 74.3                 | 32.8        | 59.3        |
| Two-stream {RGB, depth} (late fusion)             | 67.1                   | 40.9        | 48.4        | 76.5                 | 34.1        | 60.5        |
| Two-stream {RGB, texture} (early fusion)          | 65.7                   | 39.9        | 47.9        | 75.3                 | 33.3        | 59.7        |
| Two-stream {RGB, texture} (late fusion)           | 69.3                   | 42.5        | 51.1        | 78.8                 | 36.5        | 61.8        |
| Three-stream {RGB, depth, texture} (early fusion) | 67.8                   | 40.7        | 48.8        | 76.5                 | 34.9        | 60.6        |
| Three-stream {RGB, depth, texture} (late fusion)  | <b>72.3</b>            | <b>43.8</b> | <b>53.3</b> | <b>81.2</b>          | <b>38.2</b> | <b>63.5</b> |

**Table 2.** Comparison (overall accuracy in %) of our three-stream deep architecture with the baseline standard RGB stream on different indoor scene hierarchies available in SUN-397 dataset. The proposed three-stream deep architecture (late fusion) consistently improves the baseline standard RGB stream on all indoor scene hierarchies.

| Architecture                             | Shopping/dining | Workplace   | Home/hotel  | Transportation | Sports/leisure | Cultural    |
|--|-----------------|-------------|-------------|----------------|----------------|-------------|
| RGB deep stream (baseline)               | 38.4            | 46.5        | 44.3        | 56.1           | 63.8           | 43.6        |
| Ours {RGB, depth, texture} (late fusion) | <b>45.5</b>     | <b>52.5</b> | <b>54.3</b> | <b>64.7</b>    | <b>67.6</b>    | <b>51.3</b> |



**Fig. 4.** Example images from SUN-397 indoor categories where our approach provides the biggest increase (top) and the biggest decrease (bottom), compared to the baseline.

**Table 3.** Comparison (overall accuracy in %) with the state-of-the-art approaches.

| Method                       | Publication | MIT-67      | OCIS        | SUN-397     |
|------------------------------|-------------|-------------|-------------|-------------|
| Multi-scale hybrid CNNs [10] | CVPR 2016   | 86.0        | -           | 70.2        |
| DRCF-CNN [14]                | TIP 2016    | 71.8        | 32.0        | -           |
| SLSIF-CNN [8]                | TIP 2016    | 74.4        | -           | -           |
| PatchNets [29]               | TIP 2017    | 84.9        | -           | <b>71.7</b> |
| LSHybrid-CNNs [6]            | TIP 2017    | 83.8        | -           | 67.6        |
| Hybrid CNN models [31]       | TCSVT 2017  | 86.0        | -           | 70.7        |
| Spectral-CNNs [15]           | ICCV 2017   | 84.3        | -           | 67.6        |
| SCF-CNNs [13]                | MVA 2018    | 83.1        | -           | -           |
| This paper                   | -           | <b>86.4</b> | <b>45.3</b> | 69.2        |

military, law, politics, etc.) with 36 indoor scene classes. Note that some indoor scene categories are shared across different scene hierarchies. Table 2 shows the results obtained using the standard RGB and our three-stream network on the six scene hierarchies. Our approach provides significant gains of 7.1%, 6.0%, 10.0%, 3.8%, 7.5% and 7.3% on the six scene hierarchies (shopping/dining, Workplace, home/hotel, transportation, sports/leisure, and cultural), respectively. Figure 4 shows example images from different indoor scene categories from the SUN-397 dataset on which our three-stream architecture provides the biggest improvement (top) and the biggest drop (bottom), compared to the standard RGB network.

**State-of-the-Art Comparison:** State-of-the-art approaches employ very deep hybrid models pre-trained on both the ImageNet and Places datasets. Therefore, we also combine our late fusion based three-stream network, at the score/prediction level, with the very deep networks: ResNet-50 architecture [9]. Table 3 shows the comparison. Among existing methods, the works of [10, 31] provide superior performance with a mean classification accuracy of 86.0% on the MIT-67 dataset. Our approach achieves improved results compared to both these methods with a mean recognition rate of 86.4%. On the OCIS dataset, our approach significantly outperforms the existing DRCF-CNN [14] by achieving a mean accuracy of 45.3%. On the SUN-397 dataset, the best results are obtained by PatchNets [29] approach. Our approach obtains an average classification accuracy of 69.2%.

## 5 Conclusions

We introduced a three-stream deep architecture for monocular indoor scene recognition. In addition to the standard RGB, we proposed to integrate explicit estimated depth and texture streams in the deep learning architecture. We further investigated different fusion strategies to integrate the three sources of information. To the best of our knowledge, we are the first to investigate fusion strate-

gies to integrate RGB, estimated depth and texture information for monocular indoor scene recognition.

**Acknowledgement.** This work has been supported by the Academy of Finland project number 313988 *Deep neural networks in scene graph generation for perception of visual multimedia semantics* and the European Union's Horizon 2020 research and innovation programme under grant agreement No 780069 *Methods for Managing Audiovisual Data: Combining Automatic Efficiency with Human Accuracy*. Computational resources have been provided by the Aalto Science-IT project and NVIDIA Corporation.

## References

1. Anwer, R.M., Khan, F.S., van de Weijer, J., Molinier, M., Laaksonen, J.: Binary patterns encoded convolutional neural networks for texture recognition and remote sensing scene classification. *ISPRS J. Photogramm. Remote Sens.* **138**, 74–85 (2018)
2. Borg, I., Groenen, F.: *Modern Multidimensional Scaling: Theory and Applications*. Springer, New York (2005). <https://doi.org/10.1007/0-387-28981-X>
3. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets. In: *BMVC* (2014)
4. Cimpoi, M., Maji, S., Vedaldi, A.: Deep filter banks for texture recognition and segmentation. In: *CVPR*, pp. 3828–3836 (2015)
5. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: *ICCV* (2015)
6. Guo, S., Huang, W., Wang, L., Qiao, Y.: Locally supervised deep hybrid model for scene recognition. *TIP* **26**(2), 808–820 (2017)
7. Gupta, S., Arbelaez, P., Girshick, R., Malik, J.: Local binary features for texture classification: taxonomy and experimental study. *IJCV* **112**(2), 133–149 (2015)
8. Hayat, M., Khan, S., Bennamoun, M., An, S.: A spatial layout and scale invariant feature representation for indoor scene classification. *TIP* **25**(10), 4829–4841 (2016)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: *CVPR* (2016)
10. Herranz, L., Jiang, S., Li, X.: Scene recognition with CNNs: objects, scales and dataset bias. In: *CVPR* (2016)
11. Hoffman, J., Gupta, S., Darrell, T.: Learning with side information through modality hallucination. In: *CVPR* (2016)
12. Khan, F.S., Anwer, R.M., van de Weijer, J., Felsberg, M., Laaksonen, J.: Compact color-texture description for texture classification. *PRL* **51**, 16–22 (2015)
13. Khan, F.S., van de Weijer, J., Anwer, R.M., Bagdanov, A., Felsberg, M., Laaksonen, J.: Scale coding bag of deep features for human attribute and action recognition. *MVA* **29**(1), 25–71 (2018)
14. Khan, S., Hayat, M., Bennamoun, M., Togneri, R., Sohel, F.: A discriminative representation of convolutional features for indoor scene recognition. *TIP* **25**(7), 3372–3383 (2016)
15. Khan, S., Hayat, M., Porikli, F.: Scene categorization with spectral features. In: *ICCV* (2017)
16. Koskela, M., Laaksonen, J.: Convolutional network features for scene recognition. In: *ACM MM* (2014)

17. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: CVPR, pp. 2169–2178 (2006)
18. Levi, G., Hassner, T.: Emotion recognition in the wild via convolutional neural networks and mapped binary patterns. In: ICMI (2015)
19. Li, B., Shen, C., Dai, Y., van den Hengel, A., He, M.: Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs. In: CVPR (2015)
20. Liu, F., Shen, C., Lin, G., Reid, I.: Learning depth from single monocular images using deep convolutional neural fields. PAMI **38**(10), 2024–2039 (2016)
21. Liu, L., Fieguth, P., Guo, Y., Wang, X., Pietikainen, M.: Local binary features for texture classification: taxonomy and experimental study. PR **62**, 135–160 (2017)
22. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI **24**(7), 971–987 (2002)
23. Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: CVPR (2009)
24. Renninger, L.W., Malik, J.: When is scene identification just texture recognition? Vis. Res. **44**(19), 2301–2311 (2004)
25. Rubner, Y., Tomasi, C., Guibas, L.: The earth mover’s distance as a metric for image retrieval. IJCV **40**(2), 99–121 (2000)
26. Russakovsky, O., et al.: ImageNet large scale visual recognition challenge. arXiv preprint [arXiv:1409.0575](https://arxiv.org/abs/1409.0575) (2014)
27. Song, X., Herranz, L., Jiang, S.: Depth CNNs for RGB-D scene recognition: learning from scratch better than transferring from RGB-CNNs. In: AAAI (2017)
28. Wang, A., Cai, J., Lu, J., Cham, T.J.: Modality and component aware feature fusion for RGB-D scene classification. In: CVPR (2016)
29. Wang, Z., Wang, L., Wang, Y., Zhang, B., Qiao, Y.: Weakly supervised patchnets: describing and aggregating local patches for scene recognition. TIP **26**(4), 2028–2041 (2017)
30. Xiao, J., Hays, J., Ehinger, K., Oliva, A., Torralba, A.: Sun database: large-scale scene recognition from abbey to zoo. In: CVPR (2010)
31. Xie, G.S., Zhang, X.Y., Yan, S., Liu, C.L.: Hybrid CNN and dictionary-based models for scene recognition and domain adaptation. TCSVT **27**(6), 1263–1274 (2016)
32. Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., Oliva, A.: Learning deep features for scene recognition using places database. In: NIPS, pp. 487–495 (2014)