A Spatial Layout and Scale Invariant Feature Representation for Indoor Scene Classification

Munawar Hayat, Salman H. Khan, Mohammed Bennamoun and Senjian An

Abstract—**Unlike standard object classification, where the image to be classified contains one or multiple instances of the same object, indoor scene classification is quite different since the image consists of multiple distinct objects. Further, these objects can be of varying sizes and are present across numerous spatial locations in different layouts. For automatic indoor scene categorization, large-scale spatial layout deformations and scale variations are therefore two major challenges and the design of rich feature descriptors which are robust to these challenges is still an open problem. This paper introduces a new learnable feature descriptor called "spatial layout and scale invariant convolutional activations" to deal with these challenges. For this purpose, a new Convolutional Neural Network architecture is designed which incorporates a novel 'Spatially Unstructured' layer to introduce robustness against spatial layout deformations. To achieve scale invariance, we present a pyramidal image representation. For feasible training of the proposed network for images of indoor scenes, the paper proposes a methodology which efficiently adapts a trained network model (on a large-scale data) for our task with only a limited amount of available training data. The efficacy of the proposed approach is demonstrated through extensive experiments on a number of datasets including MIT-67, Scene-15, Sports-8, Graz-02 and NYU datasets.**

Index Terms—**Indoor Scenes Classification, Spatial Layout Variations, Scale Invariance**

I. INTRODUCTION

Recognition/classification is an important computer vision problem and has gained significant research attention over the last few decades. Most of the efforts, in this regard, have been tailored towards generic object recognition (which involves an image with one or multiple instances of the same object) and face recognition (which involves an image with the face region of the person).

M. Hayat is with the University of Canberra (UC), University Dr, Bruce ACT 2617, Australia. S. H. Khan is with the Data61, Commonwealth Scientific and Industrial Research Organization (CSIRO) and the Australian National University (ANU), Canberra ACT 0200, Australia. M. Bennamoun and S. An are with the University of Western Australia (UWA), Crawley, 6009, Australia. E-mail: munawar.hayat@canberra.edu.au, salman.khan@data61. csiro.au, {mohammed.bennamoun, senjian.an}@uwa.edu.au

Unlike these classification tasks, indoor scene classification is quite different since an image of an indoor scene contains multiple distinct objects, with different scales, sizes and laid across different spatial locations in a number of possible layouts. Due to the challenging nature of the problem, the state of the art performances for scene classification are much lower (*e.g.*, accuracies of $\sim 50\%$ on Places database [64] and $\sim 80\%$ on MIT-67 [4]) compared with other classification tasks such as object classification ($\sim 90\%$ mAP on PASCAL VOC-2012 [56] and $\sim 94\%$ rank-5 identification rate on ImageNet database [46]) and face recognition (*e.g.*, human-level performance on real life datasets including Labelled Faces in the Wild and YouTube Faces [50]). This paper proposes a novel method of feature description, specifically tailored for indoor scene images, in order to address the challenges of large-scale spatial layout deformations and scale variations.

We can characterize some indoor scenes by only global spatial information [33], [40], whereas for others, local appearance information [6], [22], [29] is more critical. For example, a corridor can be predominantly characterized by a single large object (*walls*) whereas a bedroom scene is characterized by multiple objects (*e.g.*, *sofa*, *bed*, *table*). Both global and local spatial information must, therefore, be leveraged in order to accommodate different scene types [39]. However, this is very challenging, for two main reasons. First, the spatial scale of the constituent objects varies significantly across different scene types. Second, the constituent objects can be present in different spatial locations and in a number of possible layouts. This is illustrated in the example images of the kitchen scene in Fig. 1, where a microwave can be present in many locations in the image with significant variations in scale, pose and appearance.

This paper aims to achieve invariance with respect to the spatial layout and the scale of the constituent objects for indoor scene images. For this purpose, in order to achieve invariance with respect to the spatial scale of objects, we generate a pyramidal image representation where an image is resized to different scales, and features are computed across these scales (Sec III-C). To achieve spatial layout invariance, we introduce a new method of

Fig. 1: The spatial structure of indoor scenes is loose, irregular and unpredictable which can confuse the classification system. As an example, a microwave in a kitchen scene can be close to the sink, fridge, kitchen door or top cupboards (*green* box in the images). Our objective is to learn feature representations which are robust to these variations by spatially shuffling the convolutional activations (Sec. III).

feature description which is based on a proposed modified Convolutional Neural Network (CNN) architecture (Sec. III-A).

CNNs preserve the global spatial layout in an image. This is desirable for the classification tasks where an image predominantly contains only a single object (*e.g.*, objects in the ImageNet database [41]). However, for a high-level vision task such as indoor scene classification, an image may contain multiple distinct objects across different spatial locations. We, therefore, want to devise a method of feature description which is robust with respect to the spatial layout of objects in a scene. Although, the local pooling layers (max or mean pooling) incorporated in standard CNN architectures do achieve viewpoint and pose invariance to some extent [20]. Further, invariance with respect to fixed sizes of input images can be achieved by employing a spatial pooling layer after the last convolutional layer as in [12]. The introduced pooling layer in [12] can generate a fixed size output which is fed into the fully connected layer, thus enabling the network to handle input images of any dimensions. These pooling layers, however, cannot accommodate large-scale deformations that are caused by spatial layout variations in indoor scenes. In order to achieve spatial layout invariance, this paper introduces a modified CNN architecture with an additional layer, termed 'spatially unstructured layer' (Sec. III-A). The proposed CNN is then trained with images of indoor scenes (using our proposed strategy described in Sec. III-B) and the learned feature representations are invariant to the spatial layout of the constituent objects.

Training a deep CNN requires a large amount of data because the number of parameters to be learned is quite huge. However, for many classification tasks, we only have a limited amount of annotated training data. This becomes then a serious limitation for the feasible training of a deep CNN. Some recently proposed techniques demonstrate that pre-trained CNN models (on large datasets *e.g.*, ImageNet) can be adapted for similar tasks with limited additional training data [3]. However, cross-domain adaptation becomes problematic in the case of heterogeneous tasks due to the different natures of source and target datasets. For example, an image in the ImageNet dataset contains mostly centred objects belonging to only one class. In contrast, an image in an indoor scene dataset has many constituent objects, all appearing in a variety of layouts and scales. In this work, we propose an efficient strategy to achieve cross-domain adaptation with only a limited number of annotated training images in the target dataset (Sec. III-B). We note that the recently released Places database [64] makes it feasible to train a large-scale CNN model for the particular case of scene classification. Our proposed approach is however quite generic and can be applied to similar problems where cross-domain adaptation of pre-trained deep CNN models is required. Specifically, while [64] introduces a new large-scale scene dataset, our approach makes it possible to perform training with only a limited amount of annotated data.

The major contributions of this paper can be summarized as: 1) A new method of feature description (using the activations of a deep convolutional neural network) is proposed to deal with the large-scale spatial layout deformations in scene images (Sec III-A), 2) A pyramidal image representation is proposed to achieve scale invariance (Sec III-C), 3) A novel transfer learning approach is introduced to efficiently adapt a pre-trained network model (on a large dataset) to any target classification task with only a small amount of available annotated training data (Sec III-B) and 4) Extensive experiments are performed to validate the proposed approach. Our results show a significant performance improvement for the challenging indoor scene classification task on a number of datasets.

II. RELATED WORK

Indoor scene classification has been actively researched and a number of methods have been developed in recent years [22], [36], [39], [40], [47], [48], [58]. While some of these methods focus on the holistic properties of scene images (*e.g.*, CENTRIST [58], Gist descriptor [33]), others give more importance to the local distinctive aspects (*e.g.*, dense SIFT [22], HOG [59]).

In this paper, we argue that we cannot rely on either of the local or holistic image characteristics to describe all indoor scene types [39]. For some scene types, holistic or global image characteristics are enough (*e.g.*, *corridor*), while for others, local image properties must be considered (*e.g.*, *bedroom*, *shop*). We therefore neither focus on the global nor the local feature description and instead extract mid-level image patches to encode an intermediate level of information. Further, we propose a pyramidal image representation which is able to capture the discriminative aspects of indoor scenes at multiple levels.

Recently, mid-level representations have emerged as a competitive candidate for indoor scene classification. Strategies have been devised to discover discriminative mid-level image patches which are then encoded by a feature descriptor. For example, the works [5], [16], [48] learn to discover discriminative patches from the training data. Our proposed method can also be categorized as a mid-level image patches based approach. However, our method is different from the previous methods, which require discriminative patch ranking and selection procedures or involve the learning of distinctive primitives. In contrast, our method achieves state of the art performance by simply extracting mid-level patches densely and uniformly from an image which makes our approach more efficient (see more details in Sec. III-D). Furthermore, similar to [46], we perform a dense patch extraction at multiple scales to take into consideration the contextual information of the scene at a number of scales. We note that similar to [38], our approach can also be used in conjunction with the patch discovery methods to further improve the classification performance (but at a relatively high computational cost).

An open problem in indoor scene classification is the design of feature descriptors which are robust to global layout deformations. The initial efforts to resolve this problem used bag-of-visual-words models or variants (*e.g.*, [1], [22], [60]), which are based on locally invariant descriptors *e.g.*, SIFT [28]. Recently, these local feature representations have been outperformed by learned feature representations from deep neural networks [11], [17], [20], [40], [41]. However, since there is no inherent mechanism in these deep networks to deal with the high variability of indoor scenes, several recent efforts have been made to fill in this gap (*e.g.*, [9], [12]). The bag of features approach of Gong et al. [9] performs VLAD pooling [14] of CNN activations. Another example is the combination of spatial pyramid matching and CNNs (proposed by He et al. [12]) to increase the feature's robustness. These methods, however, devise feature rep-

resentations on top of CNN activations and do not inherently equip the deep architectures to effectively deal with the large deformations. In contrast, this work provides an alternative strategy based on an improved network architecture to enhance invariance towards large-scale deformations. In this regard, our approach is close to the recent work of Jaderberg *et. al* [13] which introduces a differentiable feature transformer module in CNNs to learn spatially invariant feature representations. The detailed description of our proposed feature representation method is presented next.

III. Proposed Spatial Layout and Scale Invariant Convolutional Activations - S^{2}ICA

The block diagram of our proposed Spatial Layout and Scale Invariant Convolutional Activations (S^{2}ICA) based feature description method is presented in Fig 2. The detailed description of each of the blocks is given here. We first present our baseline CNN architecture followed by a detailed description of our spatially unstructured layer in Sec. III-A. Note that the spatially unstructured layer is introduced to achieve invariance to large-scale spatial deformations, which are commonly encountered in images of indoor scenes. The baseline CNN architecture is pre-trained for a large-scale classification task. A novel method is then proposed to adapt this pre-trained network for the specific task of scene categorization (Sec. III-B). Due to the data hungry nature of CNNs, it is not feasible to train a deep architecture with only a limited amount of available training data. For this purpose, we pre-train a 'TransferNet', which is then appended with the initialized CNN and the whole network can then be efficiently fine-tuned for the scene classification task. Convolutional activations from this fine-tuned network are then used for a robust feature representation of the input images. To deal with the scale variations, we propose a pyramidal image representation and combine the activations from multiple levels which result in a scale invariant feature representation (Sec. III-C). This representation is then finally used by a linear Support Vector Machine (SVM) for classification (Sec. III-D).

A. CNN Architecture

Our baseline CNN architecture is presented in Fig 4. It consists of five convolutional layers and four fully connected layers. The architecture of our baseline CNN is similar to AlexNet [20]. The main difference is that we introduce an extra fully connected layer, and that all of our neighbouring layers are densely connected (in contrast to the sparse connections in AlexNet). To achieve spatial layout invariance, the architecture of
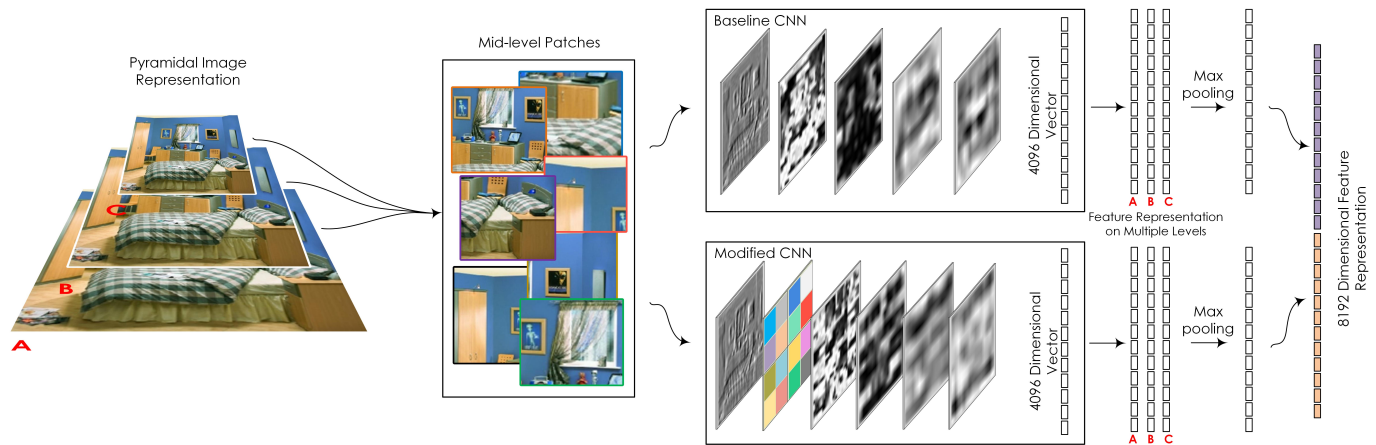
Fig. 2: Overview of the proposed Spatial Layout and Scale Invariant Convolutional Activations (S^{2}ICA) based feature description method. Mid-level patches are extracted from three levels (A, B, C) of the pyramidal image representation. The extracted patches are separately feed-forwarded to the two trained CNNs (with and without the spatially unstructured layer). The convolutional activations based feature representation of the patches is then pooled and a single feature vector for the image is finally generated by concatenating the feature vectors from both CNNs. Figure best seen in colour

the baseline CNN is modified and a new unstructured layer is added after the first sub-sampling layer. A brief description of each layer of the network follows next.

Let us suppose that the convolutional neural network consists of L hidden layers and each layer is indexed by $l \in \{1 \dots L\}$. The feed-forward pass can be described as a sequence of convolution, optional sub-sampling and normalization operations. The response of each convolution node in layer l is given by:

$$\mathbf{a}_n^l = f\left(\sum_m (\mathbf{a}_m^{l-1} * \mathbf{k}_{m,n}^l) + b_n^l\right), \qquad (1)$$

where $\mathbf{k}$ and b denote the learned kernel and bias, the indices (m,n) indicate that the mapping is from the mth feature map of the previous layer to the nth feature map of the current layer. The function f is the element-wise Rectified Linear Unit (ReLU) activation function [32]. The response of each normalization layer is given by:

$$\mathbf{a}_n^l = \frac{\mathbf{a}_n^{l-1}}{\left(\alpha + \beta \sum_{j=max(0,n-\sigma)}^{min(N-1,n+\sigma)} (\mathbf{a}_j^{l-1})^2\right)^\gamma}, \qquad (2)$$

where $\alpha, \beta, \gamma, \sigma$[1] are constants and N is the total number of kernels in the layer. The response of each sub-sampling node is given by:

$$\mathbf{a}_{n(i,j)}^l = \max_{0 \leq \{i',j'\} \leq T-1} \mathbf{a}_{n(i+i',j+j')}^{l-1}, \qquad (3)$$

[1]These constants are defined as in [20]: $\alpha = 2, \beta = 10^{-4}, \gamma = 3/4$ and $\sigma = 5/2$.

where, $\{i, i', j, j'\}$ are the input and output indexes and T is the neighbourhood size over which the values are pooled.

In our proposed modified CNN architecture, a spatially unstructured layer follows the first sub-sampling layer and breaks the spatial order of the output feature maps. This helps in the generation of robust feature representations that can cope with the high variability of indoor scenes. Algorithmic description of the operations performed by the spatially unstructured layer is presented in Algorithm 1. For each feature response, we split the feature map into a specified number of blocks (b). Next, a matrix $\mathbf{U}$ is constructed whose elements correspond to the scope of each block defined as a tuple:

$$\mathbf{U}_{\sqrt{b} \times \sqrt{b}} = \{\mathbf{u}_i \ \forall i \ |\mathbf{u}_i = (p,q)\}, \qquad (4)$$

where, p and q indicate the starting and ending index of each block. To perform a local swapping operation, we define a matrix $\mathbf{S}$ in terms of an identity matrix I as follows:

$$\mathbf{S}_{2 \times 2} = |I - 1| = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \qquad (5)$$

Next, a transformation matrix $\mathbf{T} \in \mathbb{R}^{\sqrt{b} \times \sqrt{b}}$ is defined in terms of $\mathbf{S}$ as follows:

$$\mathbf{T}_{\sqrt{b} \times \sqrt{b}} = \begin{pmatrix} \mathbf{S} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{S} & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{S} \end{pmatrix}_{\sqrt{b}/2 \times \sqrt{b}/2} \qquad (6)$$

The transformation matrix $\mathbf{T}$ has the following properties:

Algorithm 1 Operations Involved in the Spatially Unstructured Layer

---

Input: Feature map, $\mathbf{F} \in \mathbb{R}^{p \times q \times r \times s}$, Number of Blocks : N

Output: Modified feature map ($\mathbf{F}_m$)

$\ell = \left\lfloor \frac{\sqrt{N}}{2} \right\rfloor$ // Rearrangement level

$h_{pts} \leftarrow (\ell + 1)$ linearly spaced points in range $[1 : p]$

$h_{pts}[end] += 1$

$w_{pts} = h_{pts}$ // $\because p = q$ for $\mathbf{F}$

for $\forall i \in [1 : length(h_{pts}) - 1]$ **do**

    for $\forall j \in [1 : length(w_{pts}) - 1]$ **do**

        $F_{tmp} = \mathbf{F}[h_{pts}(i) : h_{pts}(i+1) - 1, w_{pts}(j) : w_{pts}(j+1) - 1, :, :]$

        $F_{tmp} = [F_{tmp}(\left\lceil \frac{rows(F_{tmp})}{2} \right\rceil : end, :, :, :); F_{tmp}(1 : \left\lfloor \frac{rows(F_{tmp})}{2} \right\rfloor, :, :, :)]$

        $F_{tmp} = [F_{tmp}(:, \left\lceil \frac{cols(F_{tmp})}{2} \right\rceil : end, :, :); F_{tmp}(:, 1 : \left\lfloor \frac{cols(F_{tmp})}{2} \right\rfloor, :, :)]$

        $\mathbf{F}_m[h_{pts}(i) : h_{pts}(i+1) - 1, w_{pts}(j) : w_{pts}(j+1) - 1, :, :] = F_{tmp}$

return $\{\mathbf{R}\}$

- $\mathbf{T} = \{t_{ij}\}$ is a permutation matrix ($\mathbf{T} : \{\mathbf{u}_{ij}\} \to \{\mathbf{u}_{ij}\}$) since the sum along each row and column is always equal to one *i.e.* , $\sum_i t_{ij} = \sum_j t_{ij} = 1$.

- $\mathbf{T}$ is a bi-stochastic matrix and therefore according to Birkhoffvon Neumann theorem and the above property, $\mathbf{T}$ lies on the convex hull of the set of bi-stochastic matrices.

- It is a binary matrix with entries belonging to the Boolean domain $\{0, 1\}$.

- It is an orthogonal matrix, therefore, $\mathbf{T}\mathbf{T}^T = I$ and $\mathbf{T}^{-1} = \mathbf{T}^T$.

Using the matrix $\mathbf{T}$, we transform $\mathbf{U}$ to become:

$$\hat{\mathbf{U}} = (\mathbf{U}^T \mathbf{T})^T \mathbf{T} = \mathbf{T}^T \mathbf{U} \mathbf{T}. \qquad (7)$$

The updated matrix $\hat{\mathbf{U}}$ contains the new indices of the modified feature maps. If $\mathcal{Y}(\cdot)$ is a function which reads the indices of the blocks stored in the form of tuples in matrix $\hat{\mathbf{U}}$, the layer outputs are as follows:

$$\mathbf{a}_n^l = r * \mathcal{Y}(\mathbf{a}_n^{l-1}, \hat{\mathbf{U}}), \qquad (8)$$

$$\text{where, } r \sim \text{Bernoulli}(\rho). \qquad (9)$$

r is a random variable which has a probability ρ of being equal to 1. Note that this shuffling operation is applied randomly so that a network does not get biased towards the normal patches. Fig. 3 illustrates the distortion operations performed by the spatially unstructured layer for a varying number of blocks.

B. Training CNNs for Indoor Scenes

Deep CNNs have demonstrated exceptional feature representation capabilities for the classification and detection tasks (*e.g.*, see ILSVRC'14 Results [41]). Training deep CNNs however requires a large amount of data
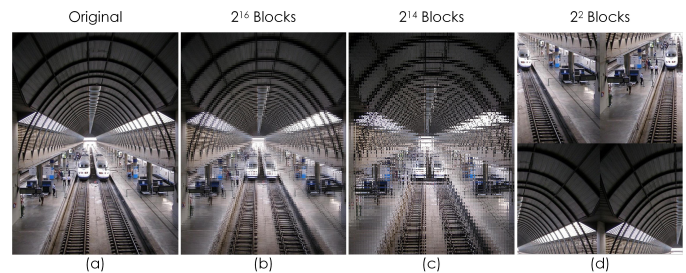

Fig. 3: (*left* to *right*) Original image and the spatially unstructured versions with 2^{16}, 2^{14} and 2^2 blocks respectively.

since the number of parameters to be learned is huge. The requirement of a large amount of training data makes the training of CNNs infeasible where only a limited amount of annotated training data is available. In this paper, we propose to leverage from the image representations learned on a large-scale classification task (such as on ImageNet [41]) and propose a strategy to learn tailored feature representations for indoor scene categorization. In this manner, our approach performs cross-domain feature adaptation by transforming the object-centric features to scene-centric feature representations. An algorithmic description of our proposed strategy is summarized in Algorithm. 2. The details are presented here.

We first train our baseline CNN architecture on ImageNet database following the procedure in [20]. Next, we densely extract mid-level image patches from our scene classification training data and represent them in terms of the convolutional activations of the trained baseline network. The output of the last convolution layer followed by ReLU non-linearity is considered as
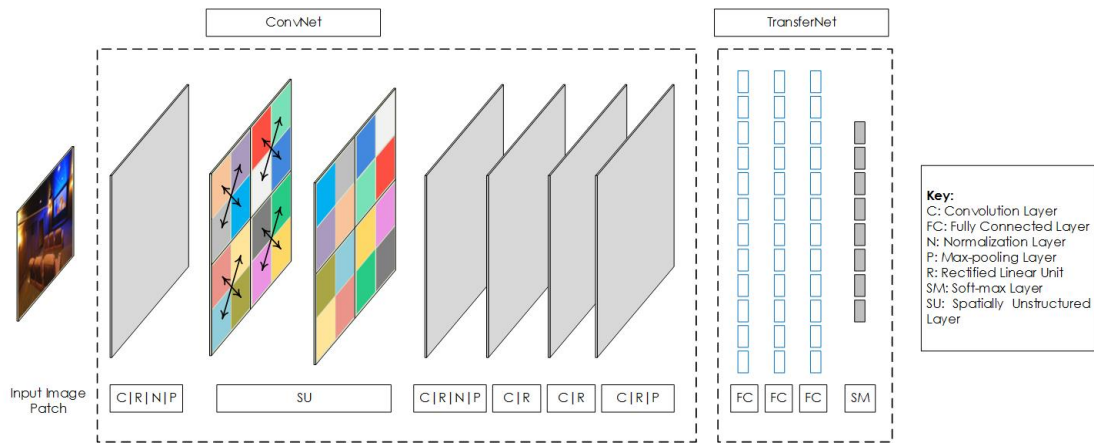
Fig. 4: The architecture of our proposed Convolutional Neural Network used to learn tailored feature representations for scene categorization. We devise a strategy (see Sec. III-B and Alg. 2) to effectively adapt the learned feature representation from a large-scale classification task to scene categorization.

a feature representation of the extracted patches. These feature representations ($\mathcal{F}$) will be used to train our TransferNet.

Algorithm 2 Training CNNs for indoor scenes

Input: Source DB (ImageNet), Target DB (Scene Images)

Output: Learned weights: $\{\mathbf{W}\}_{1 \times L}$, $\{\mathbf{W}_{su}\}_{1 \times L}$

 1: Pre-train the CNN on the large-scale source DB.
 2: Feed-forward image patches from target DB to trained CNN.
 3: Take feature representations ($\mathcal{F}$) from the last convolution layer.
 4: Train the 'TransferNet' consisting of four fully connected layers with $\mathcal{F}$ as input and target annotations as output.
 5: Append 'TransferNet' to the last convolution layer of trained CNN.
 6: Fine-tune the complete network with and without the spatially unstructured layer to get $\{\mathbf{W}\}_{1 \times L}$ and $\{\mathbf{W}_{su}\}_{1 \times L}$ respectively.

As depicted in Fig 4, our TransferNet consists of three hidden layers (with 4096 neurons each) and an output layer, whose number of neurons are equal to the number of classes in the target dataset (*e.g.*, indoor scenes dataset). TransferNet is trained on convolutional feature representations ($\mathcal{F}$) of mid-level patches of the scene classification dataset. Specifically, the input to TransferNet are the feature representations ($\mathcal{F}$) of the patches and the outputs are their corresponding class labels. After training TransferNet, we remove all fully connected layers of the baseline CNN and join the trained TransferNet to the last convolutional layer of the

baseline CNN. The resulting network then consists of five convolutional layers and four fully connected layers (of the trained TransferNet). This complete network is now fine-tuned on the patches extracted from the training images of the scene classification data. Since the network initialization is quite good (the convolutional layers of the network are initialized from the baseline network trained on ImageNet dataset, whereas the fully connected layers are initialized from the trained transferNet), only few epochs are required for the network to converge. Moreover, with a good initialization, it becomes feasible to learn deep CNN's parameters even with a smaller number of available training images.

Our proposed fine-tuning strategy is similar to [8], [35] in which a lower learning rate (zero or close to zero) is used for the convolutional part of the network (thus essentially freezing this part) and a higher learning rate is used for the fully connected part. Adapting the fine tuning strategy of [8], [35] will yield similar tuned parameters. However, a major advantage of our proposed technique is that it does not need to compute the activations for the convolutional part of the network in every feed forward pass during the training phase. Instead, these are computed only once at the beginning, which makes our proposed technique computationally efficient.

Note that the baseline CNN was trained with images from the ImageNet database, where each image predominantly contains one or multiple instances of the same object. In the case of scene categorization, we deal with a large number of object categories, where each category may appear in a variety of poses, appearances and scales at different spatial locations in a scene. Therefore, in order to incorporate large-scale deformations, we train

two CNNs: with and without the spatially unstructured layer (learned weights represented by $\mathbf{W}$ and $\mathbf{W}_{su}$ respectively). These trained CNNs are then used for the proposed feature representation in Sec. III-D. Below, we first explain our approach to deal with large-scale variations in object size.

C. Pyramid Image Representation

In order to achieve scale invariance, we generate a pyramid of an image at multiple spatial resolutions. However, unlike conventional pyramid generation processes (*e.g.*, Gaussian or Laplacian pyramid) where smoothing and sub-sampling operations are repeatedly applied, we simply resize each image to a set of scales and this may involve up or down sampling. Specifically, we transform each image to three scales, $\{0.75D, D, 1.25D\}$, where D is the smaller dimension of an image which is set based on the given dataset. At each scale, we densely extract patches which are then encoded in terms of the convolutional activations of the trained CNNs.

D. Image Representation and Classification

From each of the three images of the pyramidal image representation, we extract multiple overlapping patches of 224×224 using a sliding window. A shift of 32 pixels is used between patches. The extracted image patches are then feed forwarded to the trained CNNs (both with and without the spatially unstructured layer). The convolutional feature representation of the patches are max-pooled to get a single feature vector representation for the image. This is denoted by A, B and C corresponding to three images of the pyramid in Fig 2. We then max pool the feature representations of these images and generate one single representation of the image for each network (with and without the spatially unstructured layer). The final feature representation is achieved by concatenating these two feature vectors. After encoding the spatial layout and the scale invariant feature representations for the images, the next step is to perform classification. We use a simple linear Support Vector Machine (SVM) classifier for this purpose.

IV. Experiments and Evaluation

The proposed approach is validated through extensive experiments on a number of datasets. To this end, we perform experiments on three indoor scene datasets (MIT-67, NYU and Scene-15). Amongst these datasets, MIT-67 is the largest dataset for indoor scene classification. The dataset is quite challenging since images of many classes are similar in appearance and thus hard to classify (see Fig. 9). Apart from indoor scene classification, we further validate our approach on two other tasks i.e., event and object datasets (Graz-02 and Sports-8). For each scene dataset, we set the hyper-parameters (including p, n, the learning rates and the number of training epochs) via cross-validation on a small held-out validation set. In Sec. IV-A below, we first present a brief description about each of the datasets and adopted experimental protocols. We then present our experimental results along with a comparison with existing state of the art and baseline approaches in Sec. IV-B and Sec. IV-C, respectively. An ablative analysis to study the individual effect of each component on the proposed method is also presented in Sec. IV-E.

A. Datasets

The **MIT-67 Dataset** contains a total of 15620 images of 67 indoor scene classes. For our experiments, we follow the standard evaluation protocol in [39]. Specifically, 100 images per class are considered, out of which 80 are used for training and the remaining 20 are used for testing. We therefore have a total of 5360 and 1340 images for training and testing respectively.

The **15 Category Scene Dataset** contains images of 15 urban and natural scene classes. The number of images for each scene class in the dataset ranges from 200-400. For performance evaluation and comparison with existing state of the art, we follow the standard evaluation protocol in [22], where 100 images per class are selected for training and the rest are used for testing.

The **NYU v1 Indoor Scene Dataset** contains a total of 2347 images belonging to 7 indoor scene categories. We follow the evaluation protocol described in [45] and use the first 60% of the images of each class for training and the last 40% images for testing.

The **Inria Graz 02 Dataset** contains a total of 1096 images of three classes (bikes, cars and people). The images of this dataset exhibit a wide range of appearance variations in the form of heavy clutter, occlusions and pose changes. The evaluation protocol defined in [30] is used in our experiments. Specifically, the training and testing splits are generated by considering the first 150 odd images for training and the first 150 even images for testing.

The **UIUC Sports Event Dataset** contains 1574 images of 8 sports event categories. Following the protocol defined in [23], we used 70 and 60 randomly sampled images per category for training and testing respectively.
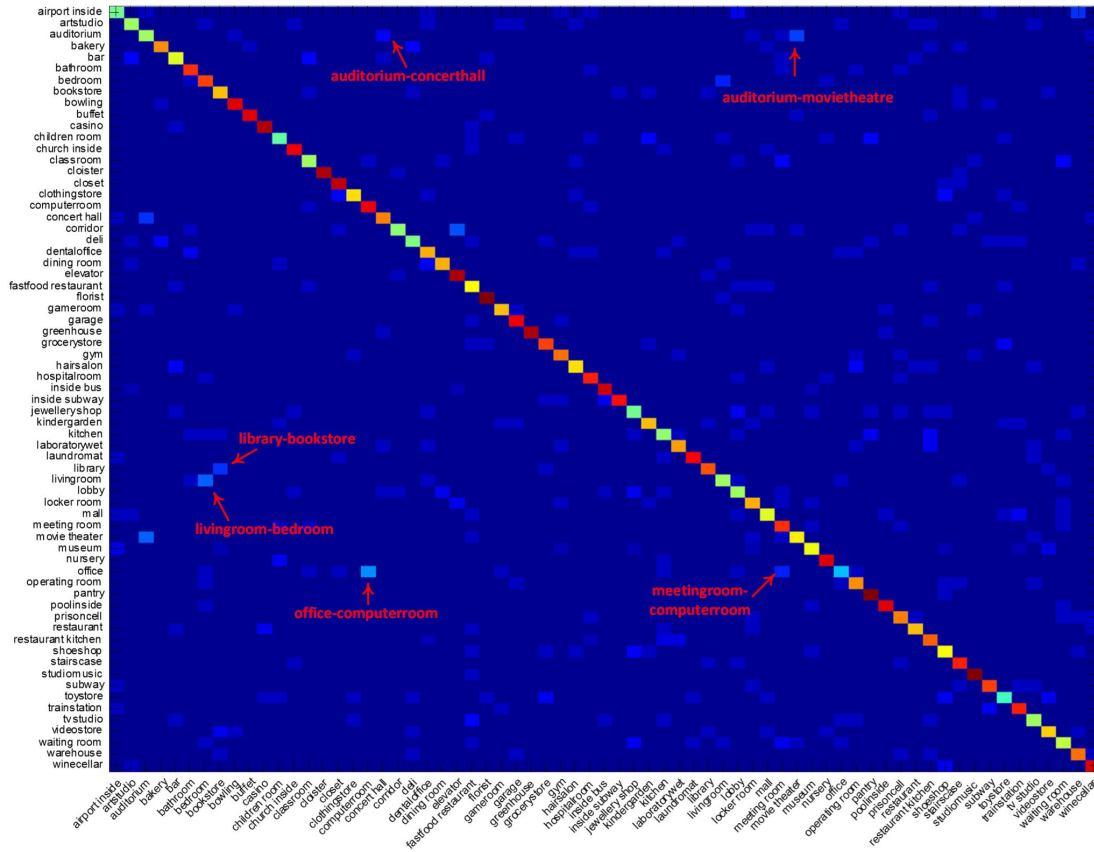
Fig. 5: Confusion Matrix for the MIT-67 Indoor Scenes Dataset. Figure best seen in color.
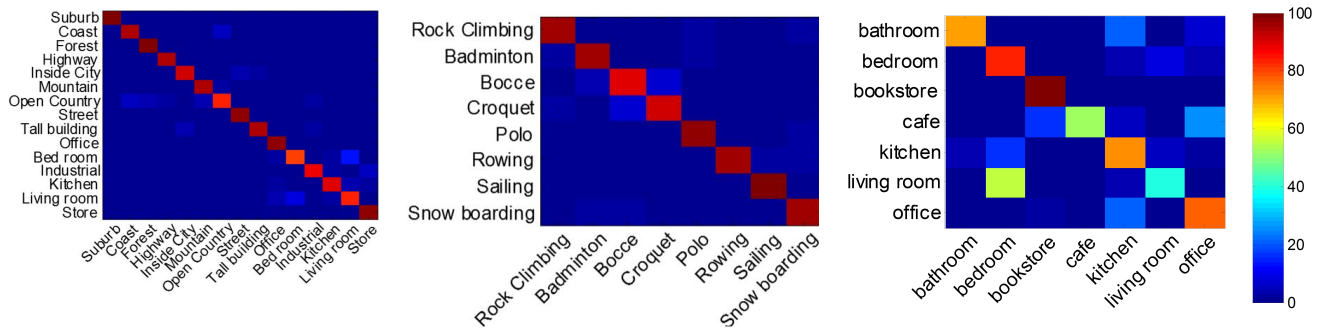

Fig. 6: Confusion matrices for Scene-15, Sports-8 and NYU scene classification datasets. Figure best seen in color.

B. Results and Analysis

The quantitative results of the proposed method in terms of classification rates for the task of indoor scene categorization are presented in Tables I, III and V. A comparison with other techniques shows that the proposed method consistently achieves a superior or at least competitive performance compared to the existing state of the art. On MIT-67 dataset, only one recent method by Cimpoi *et. al* [4] performs better than our proposed approach. However, it is important to note that [4] uses fisher vector (FV) encoding which uses high computa-

tional resources and generates high dimensional features ($\sim$ 70k compared to our $\sim$ 4k dimensional feature vector). Moreover, several recent approaches [4], [9], [18] have demonstrated that feature encoding methods, when used in conjunction with CNN activations, can further improve the classification performance. Therefore, our proposed approach can easily be extended to work with feature encoding methods to maximize its performance gain.

We also evaluate the proposed method for the tasks of sports events and highly occluded object classification

MIT-67 Indoor Scenes Dataset			
Method	Accuracy(%)	Method	Accuracy (%)
ROI + GIST [CVPR'09] [39]	26.1	ISPR [CVPR'14] [27]	50.1
MM-Scene [NIPS'10] [65]	28.3	D-Parts [ICCV'13] [48]	51.4
SPM [CVPR'06] [22]	34.4	VC + VQ [CVPR'13] [26]	52.3
Object Bank [NIPS'10] [24]	37.6	IFV [CVPR'13] [16]	60.8
RBoW [CVPR'12] [37]	37.9	MLRep [NIPS'13] [5]	64.0
Weakly Supervised DPM [ICCV'11] [36]	43.1	CNN-MOP [ECCV'14] [9]	68.9
SPMSM [ECCV'12] [21]	44.0	CNNaug-SVM [CVPRw'14] [40]	69.0
LPR-LIN [ECCV'12] [42]	44.8	Places-CNN [NIPS'14] [64]	70.8
BoP [CVPR'13] [16]	46.1	Deep Filter Banks [IJCV'16] [4]	80.3
Hybrid Parts + GIST + SP [ECCV'12] [63]	47.2		
OTC [ECCV'14] [29]	47.3	Proposed S^{2}ICA	71.2
Discriminative Patches [ECCV'12] [47]	49.4	**S^{2}ICA (with VGG-16)**	**74.4**

TABLE I: Mean accuracy on the MIT-67 indoor scenes dataset.

(Tables II and IV). The results show that the proposed method achieves very high classification rates compared to existing methods on these datasets. The overall experimental results suggest that the gain in performance of our method is more significant and pronounced for the Scene-15, Graz-02 and Sports-8 datasets. The confusion matrices showing the class-wise accuracies of Scene-15, Sports-8 and NYU datasets are presented in Fig. 6. The confusion matrix for the MIT-67 scene dataset is given in Fig. 5. It can be noted that all the confusion matrices have a very strong diagonal (Fig. 5 and 6). The majority of the confused testing samples belong to very closely related classes *e.g.*, *living room* is confused with *bedroom*, *office* with *computer-room*, *coast* with *open-country* and *croquet* with *bocce*.

The superior performance of our method is attributed to its ability to handle large spatial layout and scale variations through the introduction of the spatially unstructured layer and the proposed pyramidal image representation. Further, our method is based on deep convolutional representations, which have recently shown to be superior in performance over shallow or handcrafted feature representations [12], [40], [41]. A number of compared methods are based upon mid-level feature representations (*e.g.*, [5], [16], [48]). Our results show that our proposed method achieves superior performance over these methods. It should be noted that in contrast to existing mid-level feature representation based methods (whose main focus is on the automatic discovery of discriminative mid-level patches) our method simply densely extracts mid-level patches from uniform locations across an image. This is computationally very efficient since we do not need to devise patch selection and sorting strategies. Further, our dense patch extraction is similar to dense keypoint extraction, which has shown a comparable performance with sophisticated keypoint

UIUC Sports-8 Dataset	
Method	Accuracy (%)
GIST-color [IJCV'01] [33]	70.7
MM-Scene [NIPS'10] [65]	71.7
Graphical Model [ICCV'07] [23]	73.4
Object Bank [NIPS'10] [24]	76.3
Object Attributes [ECCV'12] [25]	77.9
CENTRIST [PAMI'11] [58]	78.2
RSP [ECCV'12] [15]	79.6
SPM [CVPR'06] [22]	81.8
SPMSM [ECCV'12] [21]	83.0
Classemes [ECCV'10] [52]	84.2
HIK [ICCV'09] [57]	84.2
LScSPM [CVPR'10] [7]	85.3
LPR-RBF [ECCV'12] [42]	86.2
Hybrid Parts + GIST + SP [ECCV'12] [63]	87.2
LCSR [CVPR'12] [44]	87.2
VC + VQ [CVPR'13] [26]	88.4
IFV [55]	90.8
ISPR [CVPR'14] [27]	89.5
Places-CNN [NIPS'14] [64]	94.2
Proposed S^{2}ICA	**95.8**

TABLE II: Mean accuracy on the UIUC Sports-8 dataset.

extraction methods over a number of classification tasks [10].

Visualisation: The contributions of the extracted mid-level patches towards a correct classification are shown in the form of heat maps for some example images in Fig 8. It can be seen that our proposed spatial layout and scale invariant convolutional activations based feature descriptor gives automatically more importance to the meaningful and information rich parts of an image.

We also study the discriminative capabilities of our proposed features by visualizing them in the original image space. For this purpose, we first embed our feature descriptors into a low dimensional 2-D space using t-SNE [54]. The embedding is done such that the original

NYU Indoor Scenes Dataset

Method	Accuracy (%)
BoW-SIFT [ICCVw'11] [45]	55.2
RGB-LLC [TC'13] [51]	78.1
RGB-LLC-RPSL [TC'13] [51]	79.5
Proposed S^{2}ICA	**81.2**
S^{2}ICA (with Places-205)	81.4

TABLE III: Mean accuracy for the NYU v1 dataset.

Graz-02 Dataset

	Cars	People	Bikes	Overall
OLB [SCIA'05] [34]	70.7	81.0	76.5	76.1
VQ [ICCV'07] [53]	80.2	85.2	89.5	85.0
ERC-F [PAMI'08] [31]	79.9	-	84.4	82.1
TSD-IB [BMVC'11] [19]	87.5	85.3	91.2	88.0
TSD-k [BMVC'11] [19]	84.8	87.3	90.7	87.6
Proposed S^{2}ICA	98.7	97.7	97.7	**98.0**

TABLE IV: Equal Error Rates (EER) on Graz-02 dataset. All performances are reported in percentages (%).

Fig. 8: The contributions (*red*: most; *blue*: least) of mid-level patches towards correct class prediction. Best seen in color.

pair-wise distances are preserved. The embedded feature descriptors in the 2-D space can then be visualized by plotting their corresponding images onto a rectangular grid. The results presented in Fig. 7 show that similar looking images lie close to each others in our proposed feature space, which is desirable for correct classification.

Ambiguous Cases: The actual and predicted labels of some miss-classified images from MIT-67 dataset are shown in Fig 9. Note the extremely challenging nature of the images in the presence of high inter-class similarities. Some of the classes are very challenging and there is no visual indication to determine the actual label. It can be seen that the miss-classified images belong to highly confusing and very similar looking scene types. For example, the image of *inside subway* is miss-classified as *inside bus*, *library* as *bookstore*, *movie theatre* as *auditorium* and *office* as *classroom*.

C. Baseline Comparisons

We attribute the superior performance of our proposed method to the following reasons: (a) its ability to handle large scale spatial layout deformations in indoor scene images (b) its ability to extract and encode information at multiple spatial levels, and (c) it provides a distinctive task specific feature representation in terms of activations of the CNN models with fine tuned fully connected parts. In order to demonstrate the effectiveness of our proposed approach to achieve large scale spatial layout and scale
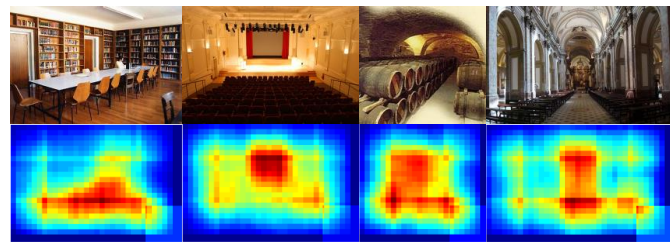
invariance, we conduct a performance comparison with several baseline methods. In this regard, we first consider standard pooling strategies in CNN models as a baseline to achieve translation invariance. We then consider activations from different layers of the CNN model as a baseline for feature encoding at different spatial levels. These baseline evaluations are discussed next.

i) Pooling for Spatial Layout Invariance: In standard CNN models, the pooling layers are quite effective and achieve a moderate level of invariance with respect to image transformations. Mean-pooling and max-pooling are the two most commonly used pooling strategies. Existing literature (*e.g.*, [43]) shows that, depending on the nature of the data, either of them can achieve better performance. Yu *et. al* [61] showed that a mixed pooling strategy, which stochastically determines the pooling type during the training of the network (instead of a pre-defined pooling type), achieves the best performance. In this baseline experiment, we study the effectiveness of different pooling strategies to achieve invariance with respect to image transformations for the task of indoor scene categorization. For this purpose, we first separately consider the activations of two CNN models (which respectively deploy max and mean pooling) as feature representation of the input image. We then concatenate these activations from both CNN models and consider the resulting vector as a feature representation. Our experimental results on MIT-67 dataset (Table VI) show that the joint feature representation from the two CNNs (one with max-pooling and the other with mean-pooling) does achieve a performance gain over the representation from either of these two models. However, in comparison to the different pooling strategies, our proposed strategy to achieve invariance with respect to spatial layout deformations (by combining complementary feature representations from the structured and unstructured CNNs) shows a significantly superior performance.

ii) Multi Layered Activations for Scale Invariance: Different layers of CNN capture different levels of in-

Fig. 7: Visualization of proposed feature representations in image space. Example images are shown from MIT-67 dataset, (a) shows examples of test images, (b) shows examples of training images. Note that images belonging to the same classes lie close to each others, demonstrating that the original inter-image distances are preserved in our proposed feature space. (Figure best seen when enlarged)
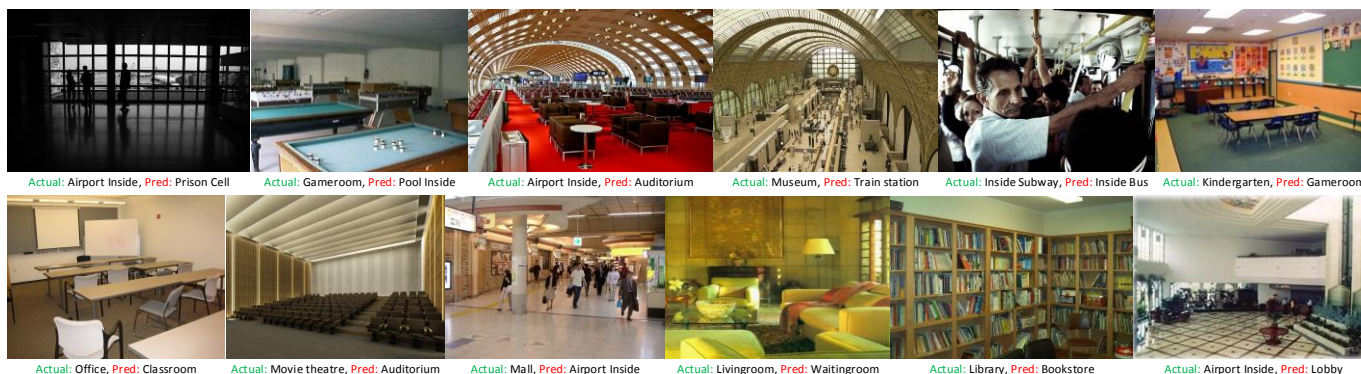

(a)

(b)

Fig. 9: Some examples of misclassified images from MIT-67 indoor scenes dataset. Actual and predicted labels of each image are given. Images from highly similar looking classes are confused amongst each other. For example, the proposed method misclassifies *library* as *bookstore*, *office* as *classroom* and *inside subway* as *inside bus*.

formation. The initial layers capture low level local geometric information while the (deeper) final layers capture more of the high level holistic global information. A possible way to simultaneously encode information at different spatial levels (local as well as global) is to combine the activations extracted from the different layers of the network. In this baseline experiment, we first individually consider activations of different layers of the CNN model as a feature representation of the input image. We then consider a joint concatenated feature representation from these layers. Our experimental results on MIT-67 datasaet (presented in Table VII) show that the best performance is achieved by combining information from the last convolution layer and the first fully connected layer. This performance, however, is quite comparable to the one achieved by using the activations of only the first fully connected layer as a feature representation. This is because the

15 Category Scene Dataset			
Method	Accuracy(%)	Method	Accuracy (%)
GIST-color [IJCV'01] [33]	69.5	ISPR [CVPR'14] [27]	85.1
Reconfigurable BoW [CVPR'12] [37]	78.6	VC + VQ [CVPR'13] [26]	85.4
Classemes [ECCV'10] [52]	80.6	LMLF [CVPR'10] [2]	85.6
Object Bank [NIPS'10] [24]	80.9	LPR-RBF [ECCV'12] [42]	85.8
SPM [CVPR'06] [22]	81.4	Hybrid Parts + GIST + SP [ECCV'12] [63]	86.3
SPMSM [ECCV'12] [21]	82.3	CENTRIST+LCC+Boosting [CVPR'11] [62]	87.8
LCSR [CVPR'12] [44]	82.7	RSP [ECCV'12] [15]	88.1
SP-pLSA [PAMI'08] [1]	83.7	IFV [55]	89.2
CENTRIST [PAMI'11] [58]	83.9	LScSPM [CVPR'10] [7]	89.7
HIK [ICCV'09] [57]	84.1	Places-CNN [NIPS'14] [64]	91.6
OTC [ECCV'14] [29]	84.4	**Proposed S^{2}ICA**	**93.1**
		S^{2}ICA (with Places-205)	**92.6**

TABLE V: Mean accuracy on the 15 Category scene dataset. Comparisons with the previous best techniques are also shown.

TABLE VI: Baseline performance evaluation of different pooling strategies on MIT-67 dataset.

Method	Accuracy (%)
Max Pooling	65.4
Mean Pooling	62.9
Max + Mean Pooling	66.5
Our Approach	**71.2**

fully connected part of the network has learned more distinctive and discriminative representations since its parameters have been adapted with respect to the specific task of indoor scenes.

TABLE VII: Performance evaluation in terms of activations of the different layers of the network as a feature representation of the input image.

Features	Accuracy (%)
Last Convolution Layer	60.2
2nd Last Convolution Layer	57.8
Last + 2nd Last Convolution Layer	60.9
First Fully Connected Layer	65.4
First Fully Connected + Last Convolution Layer	**65.6**

D. Evaluation of Baseline CNN Models

In this section, we evaluate our proposed strategies in conjunction with a range of state-of-the-art baseline CNN models. In particular, we use the VGG-16 [46], GoogleNet [49], Places-205 [64] and Places-205 Hybrid [64] models. All of these standard models have demonstrated excellent performances on large-scale classification tasks on ImageNet and Places databases. Our experimental results on MIT-67 dataset are presented

in Table VIII. The results show that a consistent performance boost is achieved over the baseline models when our proposed strategies are used to achieve spatial layout and scale invariance. We note that our approach achieves a higher performance gain on CNN models which are pre-trained on object-centric datasets (*e.g.*, VGG-16 and GoogleNet trained on ImageNet). This demonstrates its ability to adapt the object-centric feature representations that are suitable for scene-centric tasks (indoor scene classification in our case). For the case of CNN models pre-trained on scene-centric datasets (*e.g.*, Places-205), our approach achieves a relatively lower boost in performance. This can be attributed to the fact that a network trained on 2.5 million scene images (or 3.5 million for the Hybrid dataset) has already seen a large number of spatial layouts with different configurations of constituent object categories [64].

TABLE VIII: Evaluation our approach in combination with different CNN Models on the MIT-67 Dataset. All performances are reported in percentages (%).

CNN Model	Standard Model	Our Method
VGG-16 [46]	68.1	**74.4**
GoogleNet [49]	65.4	73.1
Places-205 [64]	68.2	70.2
Places-205 Hybrid [64]	**70.2**	71.5

We evaluate our proposed approach to achieve spatial layout and scale invariance on Places 2015 dataset [64]. In this regards, we consider the Places-205 CNN model and refine its parameters by incorporating the proposed spatially unstructured layer. We achieve a classification accuracy of 51.3% compared to an accuracy of 50.0% achieved with their baseline model. This validates the efficacy of our proposed strategies to achieve spatial

layout and scale invariance on a large scale dataset.

E. Ablative Analysis

An ablative analysis to assess the effect of each individual component of the proposed technique towards the overall performance is presented in Table X. Specifically, the contributions of the proposed spatially unstructured layer, pyramid image representation, training of the CNN on the target dataset and pooling (mean pooling and max pooling) are investigated. In order to investigate a specific component of the proposed method, we only modify (add or remove) that part, while the rest of the pipeline is kept fixed. The experimental results in Table X show that the combination of feature representations from CNNs trained with and without the spatially unstructured layer achieve the best performance. Furthermore, the proposed pyramidal image representation also contributes significantly towards the performance improvement of the proposed method. Our proposed strategy to adapt a deep CNN (trained on a large-scale classification task) for scene categorization also proves to be very effective and it results in a significant performance improvement. Amongst the pooling strategies, max pooling provides a superior performance compared with mean pooling.

It is interesting to note that the spatially unstructured layer disturbs the spatial structure of a scene and it helps to achieve invariance with respect to large scale spatial layout deformations. In some cases, it is not desirable to loose the spatial structure altogether (*e.g.*, a gallery scene with only a few object classes such as walls, floor and ceiling). While for some other cases, it is desirable to disturb the overall scene structure and train the network to deal with large deformations (*e.g.*, an indoor kitchen scene with hundreds of objects and cluttered regions). The combination of information from the two CNN models (structured CNN and unstructured CNN) therefore complement each other and achieves the best performance (71.2% compared to 65.4% and 65.9% of the baseline and modified CNNs respectively).

In our pyramid image representation, an image is rescaled to $\{0.75D, D, 1.25D\}$ to extract multiple mid-level patches. To analyse the impact of scales on performance, we perform experiments on MIT-67 dataset using different scales of our pyramid image representation. Our results presented in Table IX show that a combination of five scales performs slightly better than three scales. This, however, requires more computational effort. We therefore opt to use three scales considering the achieved performance and the required computational load.

Spatial Scales	Levels	Performance (%)
$0.75D, D$	2	70.1
$D, 1.25D$	2	69.8
$0.75D, D, 1.25D$	3	71.2
$0.5D, 0.75D, D, 1.25D, 1.5D$	5	**71.4**

TABLE IX: Performance evaluation for different choices of spatial scales in our pyramid image representation

Variants of Our Approach	Accuracy (%)
Baseline CNN (w/o Spatially Unstructured layer)	65.4
Modified CNN (with Spatially Unstructured layer)	65.9
Mean-pooling	65.7
w/o pyramidal representation	68.5
CNN trained on imageNet	67.3
Proposed S^{2}ICA	71.2

TABLE X: Ablative analysis on MIT-67 dataset. The joint feature representations from baseline and modified CNNs gives the best performance. The proposed pyramidal image representation results in a significant performance boost.

V. CONCLUSION

This paper proposed a novel approach to handle the large-scale deformations caused by spatial layout and scale variations in indoor scenes. A pyramidal image representation has been contrived to deal with scale variations. A modified CNN architecture with a spatially unstructured layer has been introduced to deal with the variations caused by spatial layout changes. In order to feasibly train a CNN on tasks with only a limited annotated training dataset, the paper proposed an efficient strategy which conveniently transfers learning from a large-scale dataset to a different yet related task. A robust feature representation of an image is then achieved by extracting mid-level patches and encoding them in terms of the convolutional activations of the trained networks. Leveraging on the proposed spatial layout and scale invariant image representation, the state of the art classification performance has been achieved by using a simple linear SVM classifier.

Munawar Hayat received the B.E. degree from the National University of Science and Technology, Pakistan, in 2009, and the Ph.D. degree from The University of Western Australia (UWA), in 2015. He held a post-doctoral position with IBM Research Australia. He is currently working at University of Canberra as an Assistant Professor. His research interests include computer vision, signal and image processing, pattern recognition, and machine learning. He received the Erasmus Mundus Scholarship for a joint European masters degree program. His Ph.D. thesis received the prestigious Robert Street Prize and the Deans List Honorable Mention Award at UWA.

Salman H. Khan received the B.E. degree in electrical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2012, and the Ph.D. degree from The University of Western Australia (UWA), in 2016. From 2010 to 2013, he was with research groups at the NUST and the FAST-National University of Computer and Emerging Sciences, Islamabad. He has also been a Visiting Researcher with National ICT Australia, CRL, during the year 2015, and a Researcher with Data61 (CSIRO) and Australian National University (ANU) since 2016. His research interests include computer vision, pattern recognition and machine learning.

Mohammed Bennamoun received the M.Sc. degree in control theory from Queens University, Kingston, Canada, and the Ph.D. degree in computer vision from Queens University/Queensland University of Technology, Brisbane, Australia. He is currently a Winthrop Professor with The University of Western Australia, Australia. He has authored over 300 journal and conference publications. His research interests include control theory, robotics, object recognition, artificial neural networks, signal/ image processing, and computer vision. He was selected to give conference tutorials at the European Conference on Computer Vision (ECCV) and the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). He organized several special sessions for conferences, e.g., the IEEE International Conference in Image Processing (ICIP) and the International Conference on Computer Vision (ICCV). He also contributed in the organization of many local and international conferences. He served as a Guest Editor of a couple of special issues in International journals, such as the International Journal of Pattern Recognition and Artificial Intelligence.

Senjian An received his B.S degree from Shandong University, the M.S. degree from the Chinese Academy of Sciences, and the Ph.D. degree from Peking University, China. He is currently a Research Assistant Professor at the School of Computer Science and Software Engineering, The University of Western Australia. His research interests include machine learning, image processing, object detection and recognition.

REFERENCES

[1] A. Bosch, A. Zisserman, and X. Muoz, "Scene classification using a hybrid generative/discriminative approach," *PAMI*, vol. 30, no. 4, pp. 712–727, 2008.

[2] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *CVPR*. IEEE, 2010, pp. 2559–2566.

[3] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," in *BMVC*, 2014.

[4] M. Cimpoi, S. Maji, and A. Vedaldi, "Deep filter banks for texture recognition and segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3828–3836.

[5] C. Doersch, A. Gupta, and A. A. Efros, "Mid-level visual element discovery as discriminative mode seeking," in *NIPS*, 2013, pp. 494–502.

[6] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, vol. 2. IEEE, 2005, pp. 524–531.

[7] S. Gao, I. W. Tsang, L.-T. Chia, and P. Zhao, "Local features are not lonely–laplacian sparse coding for image classification," in *CVPR*. IEEE, 2010, pp. 3555–3561.

[8] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[9] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 392–407.

[10] M. Hayat, M. Bennamoun, and A. El-Sallam, "Evaluation of spatiotemporal detectors and descriptors for facial expression recognition," in *Human System Interactions (HSI), 2012 5th International Conference on*, June 2012, pp. 43–47.

[11] M. Hayat, M. Bennamoun, and S. An, "Deep reconstruction models for image set classification," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 4, pp. 713–727, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," in *ECCV*. Springer, 2014, pp. 346–361.

[13] M. Jaderberg, K. Simonyan, A. Zisserman *et al.*, "Spatial transformer networks," in *Advances in Neural Information Processing Systems*, 2015, pp. 2008–2016.

[14] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *CVPR*. IEEE, 2010, pp. 3304–3311.

[15] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *ECCV*. Springer, 2012, pp. 730–743.

[16] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *CVPR*. IEEE, 2013, pp. 923–930.

[17] S. H. Khan, M. Bennamoun, F. Sohel, and R. Togneri, "Automatic shadow detection and removal from a single image," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 3, pp. 431–446, 2016.

[18] S. H. Khan, M. Hayat, M. Bennamoun, R. Togneri, and F. A. Sohel, "A discriminative representation of convolutional features for indoor scene recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 7, pp. 3372–3383, 2016.

[19] J. Krapac, J. Verbeek, F. Jurie *et al.*, "Learning tree-structured descriptor quantizers for image categorization," in *BMVC*, 2011.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097–1105.

[21] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *ECCV*.   Springer, 2012, pp. 359–372.

[22] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, vol. 2.   IEEE, 2006, pp. 2169–2178.

[23] L.-J. Li and L. Fei-Fei, "What, where and who? classifying events by scene and object recognition," in *ICCV*.   IEEE, 2007, pp. 1–8.

[24] 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 *NIPS*, 2010, pp. 1378–1386.

[25] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Objects as attributes for scene classification," in *Trends and Topics in Computer Vision*. Springer, 2012, pp. 57–69.

[26] Q. Li, J. Wu, and Z. Tu, "Harvesting mid-level visual concepts from large-scale internet images," in *CVPR*.   IEEE, 2013, pp. 851–858.

[27] D. Lin, C. Lu, R. Liao, and J. Jia, "Learning important spatial pooling regions for scene classification," 2014.

[28] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[29] R. Margolin, L. Zelnik-Manor, and A. Tal, "Otc: A novel local descriptor for scene classification," in *ECCV*.   Springer, 2014, pp. 377–391.

[30] M. Marszatek and C. Schmid, "Accurate object localization with shape masks," in *CVPR*.   IEEE, 2007, pp. 1–8.

[31] F. Moosmann, E. Nowak, and F. Jurie, "Randomized clustering forests for image classification," *PAMI*, vol. 30, no. 9, pp. 1632–1646, 2008.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[33] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[34] A. Opelt and A. Pinz, "Object localization with boosting and weak supervision for generic object recognition," in *SCIA*. Springer, 2005, pp. 862–871.

[35] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Learning and transferring mid-level image representations using convolutional neural networks," in *CVPR*, 2014, pp. 1717–1724.

[36] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *ICCV*.   IEEE, 2011, pp. 1307–1314.

[37] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb, "Reconfigurable models for scene recognition," in *CVPR*.   IEEE, 2012, pp. 2775–2782.

[38] S. N. Parizi, A. Vedaldi, A. Zisserman, and P. Felzenszwalb, "Automatic discovery and optimization of parts for image classification," in *ICLR*, 2015.

[39] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *CVPR*.   IEEE, 2009.

[40] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," *arXiv preprint arXiv:1403.6382*, 2014.

[41] 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," 2014.

[42] F. Sadeghi and M. F. Tappen, "Latent pyramidal regions for recognizing scenes," in *ECCV*.   Springer, 2012, pp. 228–241.

[43] D. Scherer, A. Müller, and S. Behnke, "Evaluation of pooling operations in convolutional architectures for object recognition," in *Artificial Neural Networks–ICANN 2010*.   Springer, 2010, pp. 92–101.

[44] A. Shabou and H. LeBorgne, "Locality-constrained and spatially regularized coding for scene categorization," in *CVPR*. IEEE, 2012, pp. 3618–3625.

[45] N. Silberman and R. Fergus, "Indoor scene segmentation using a structured light sensor," in *ICCVw*, 2011.

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

[47] S. Singh, A. Gupta, and A. A. Efros, "Unsupervised discovery of mid-level discriminative patches," in *ECCV*.   Springer, 2012, pp. 73–86.

[48] J. Sun and J. Ponce, "Learning discriminative part detectors for image classification and cosegmentation," in *ICCV*.   IEEE, 2013, pp. 3400–3407.

[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*.   IEEE, 2015, pp. 1–9.

[50] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *CVPR*.   IEEE, 2014, pp. 1701–1708.

[51] D. Tao, L. Jin, Z. Yang, and X. Li, "Rank preserving sparse learning for kinect based scene classification." *IEEE transactions on cybernetics*, vol. 43, no. 5, p. 1406, 2013.

[52] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *ECCV*.   Springer, 2010, pp. 776–789.

[53] T. Tuytelaars and C. Schmid, "Vector quantizing feature space with a regular lattice," in *ICCV*.   IEEE, 2007, pp. 1–8.

[54] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. 2579-2605, p. 85, 2008.

[55] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[56] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "Cnn: Single-label to multi-label," *arXiv preprint arXiv:1406.5726*, 2014.

[57] J. Wu and J. M. Rehg, "Beyond the euclidean distance: Creating effective visual codebooks using the histogram intersection kernel," in *ICCV*.   IEEE, 2009, pp. 630–637.

[58] ——, "Centrist: A visual descriptor for scene categorization," *PAMI*, vol. 33, no. 8, pp. 1489–1501, 2011.

[59] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *CVPR*.   IEEE, 2010, pp. 3485–3492.

[60] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *CVPR*.   IEEE, 2009, pp. 1794–1801.

[61] D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *Rough Sets and Knowledge Technology*.   Springer, 2014, pp. 364–375.

[62] J. Yuan, M. Yang, and Y. Wu, "Mining discriminative co-occurrence patterns for visual recognition," in *CVPR*.   IEEE, 2011, pp. 2777–2784.

[63] Y. Zheng, Y.-G. Jiang, and X. Xue, "Learning hybrid part filters for scene recognition," in *ECCV*.   Springer, 2012, pp. 172–185.

[64] 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*, 2014, pp. 487–495.

[65] J. Zhu, L.-J. Li, L. Fei-Fei, and E. P. Xing, "Large margin learning of upstream scene understanding models," in *NIPS*, 2010, pp. 2586–2594.