Multi-scale multi-feature context modeling for scene recognition in the semantic manifold

Xinhang Song, *Student Member, IEEE*, Shuqiang Jiang*, *Senior Member, IEEE*, Luis Herranz, *Member, IEEE*

***Abstract*—Before the big data era, scene recognition was often approached with two-step inference using localized intermediate representations (objects, topics, etc). One of such approaches is the semantic manifold (SM), in which patches and images are modeled as points in a semantic probability simplex. Patch models are learned resorting to weak supervision via image labels, which leads to the problem of scene categories co-occurring in this semantic space. Fortunately, each category has its own co-occurrence patterns that are consistent across the images in that category. Thus, discovering and modeling these patterns is critical to improve the recognition performance in this representation. Since the emergence of large datasets, such as ImageNet and Places, these approaches have been relegated in favor of the much more powerful convolutional neural networks (CNNs), which can automatically learn multi-layered representations from the data. In this paper we address many limitations of the original SM approach and related works. We propose discriminative patch representations using neural networks and further propose a hybrid architecture in which the semantic manifold is built on top of multiscale CNNs. Both representations can be computed significantly faster than the Gaussian mixture models of the original SM. To combine multiple scales, spatial relations and multiple features we formulate rich context models using Markov random fields. To solve the optimization problem we analyze global and local approaches, where a top-down hierarchical algorithm has the best performance. Experimental results show that exploiting different types of contextual relations jointly consistently improves the recognition accuracy.**

***Index Terms*—scene recognition, semantic manifold, semantic multinomial, multi-scale, context model, Markov random field, convolutional neural networks**

I. Introduction

SCENES (e.g. *coast, mountain, office*) are abstract semantic entities composed of many less abstract and localized ones (e.g. *sky, rock, table, car*). Accurate scene recognition is challenging because it implies reasoning from low-level visual features to high-level scene categories. While scene categories can be modeled directly using descriptors specific for scenes (e.g. GIST [1], CENTRIST [2]), this large semantic gap makes difficult to discriminate between a large number of scene categories.

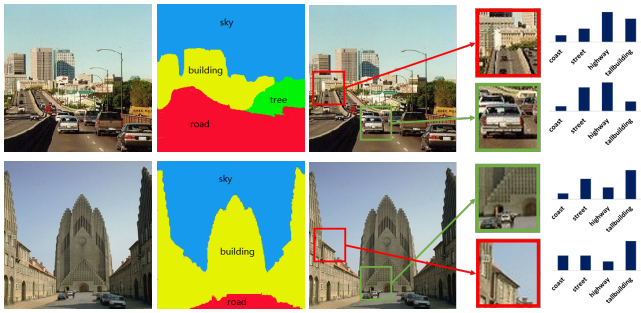


Fig. 1. Scene category co-occurrences in scene recognition: (a) images from the *highway* (top row) and *tallbuilding* (bottom row) categories of the *15 scenes* dataset, (b) regions labeled with a mid-level vocabulary, (c) patches labeled and their corresponding semantic multinomial resulting from weakly-supervised learning with scene labels. Note how the isolated patches have similar content (e.g., road, walls, cars), which introduces ambiguity and uncertainty in the estimated scene category (shown in the semantic multinomial descriptor).

A more common approach is to split the reasoning in two (or more) steps with smaller semantic gaps (e.g., features to objects, objects to scenes) [3], [4]. Thus, a local intermediate representation is defined over a vocabulary of mid-level concepts or themes. Figure 1a-b shows an example of two images and their regions with the corresponding mid-level concepts. To avoid explicitly annotating regions with mid-level labels, some approaches use latent representations, such as topic models [5], [6], [7] and discriminative parts [8], [9], [10], but the challenge is now to discover them while learning both models jointly.

The *semantic manifold* (SM) [11] uses an intermediate representation based on the *semantic multinomial* (SMN) [12], in which patches are also represented in terms of scene categories (i.e., patches are no longer represented with mid-level concepts, such as *sky*, *road*, *trees*, but scenes, such as *coast*, *street*, see Figure 1c). Patch models are learned in a weakly-supervised way using only image labels (i.e., the scene category), and thus bypassing the problems of mid-level annotations and discovering latent representations. However, this weak supervision creates a specific problem of ambiguity

This work was supported in part by the National Natural Science Foundation of China under Grant 61532018 and 61322212, in part by the Beijing Municipal Commission of Science and Technology under Grant D161100001816001, in part by the Lenovo Outstanding Young Scientists Program, in part by National Program for Special Support of Eminent Professionals and National Program for Support of Top-notch Young Professionals.

X. Song and S. Jiang are with Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China, are also with University of Chinese Academy of Sciences, Beijing, 100049, China, e-mail: {xinhang.song, shuqiang.jiang}@vipl.ict.ac.cn.

L. Herranz is with Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China, email: luis.herranz@vipl.ict.ac.cn.

*Corresponding author: Shuqiang Jiang.

which we will refer to as *(scene) category co-occurrences*[1] [13]. Fortunately, these category co-occurrences appear in patterns that are consistent across the images in the same category, so they can be modeled and separated from accidental co-occurrences (i.e., noise in the semantic representation) with an additional classifier. The way of aggregating patch SMNs into image SMNs is also tricky, since it should emphasize consistent category co-occurrence patterns so they can be modeled robustly [13], [11].

However, the SM framework still has several limitations. In particular, previous works [11], [13] only model global co-occurrences in image SMNs, while category co-occurrence patterns also appear locally (see Figure 1c). In this paper we focus on local co-occurrence patterns in patch SMNs. Our motivation is to exploit (in an unsupervised way) contextual relations to reinforce consistent co-occurrence patterns and remove accidental ones (i.e., noise). In this way, the classifier can learn a more robust model from cleaner SMN descriptors.

A second limitation of current SM frameworks is the SMN representation itself, based on GMMs. Patch SMN models are learned independently for each category, which makes them not very discriminative. At the same time, they do not scale well to datasets with many categories. Here we propose using discriminative SMN representations based on neural networks, which are learned for all categories jointly, and since they share intermediate layers they can scale much more easily to large number of categories.

In particular the contributions of this paper are:

- Analysis of the limitations of the original semantic manifold framework and its variants (Section III).
- Neural network-based discriminative SMNs, to address the problem of efficiency and lack of discriminative capability of the original GMM-SMNs (Section IV).
- Context models to exploit spatial, multi-feature and multi-scale relations between SMNs, with the objective of emphasizing consistent scene co-occurrence patterns and removing accidental ones. We formulate it in a Markov random field (MRF) framework, analyzing different context models and ways to solve the optimization problem (Section V).

The research in this paper is an extension of our previous work [14], which focuses on exploiting spatial and multi-feature context on GMM-SMNs. However, in this paper we significantly extend that framework including discriminative SMNs, (both shallow and deep architectures), multi-scale context and a hierarchical message passing algorithm to solve the optimization problem. We also include more detailed analysis of the limitations of previous works and extended experiments that achieve state-of-the-art scene recognition performance.

[1]In [13], the authors use the term *contextual co-occurrences* to refer to consistent and thus desirable co-occurrence patterns. Here, we refer to them as *(scene) category co-occurrences* to emphasize that they are high-level categories rather than low or mid-level co-occurrences. We also want to avoid confusion with other type of context, such as the spatial neighborhood, multi-scale or inter-feature relations.

II. Related work

For convenience we include Table I with several abbreviations used across the paper and the related references.

TABLE I
ABBREVIATIONS USED IN THE PAPER.

SMN	Semantic multinomial[12]	GMM	Gaussian mixture model
SM	Semantic manifold[11]	SFV	Semantic fisher vector[15]
ICM	Iterated conditional modes[16]	MP	Message passing
IM	ImageNet dataset[17]	PL	Places dataset[18]
MF	Multi-feature context[14]	EMFS	Extended MFS[14]
MFS	Multi-feature spatial context[14]	NN	Neural network
KCNF	Kernel contextual noise filter[19]	LDA	Latent Dirichlet allocation
SPM	Spatial pyramid matching[20]	CMN	Contextual multinomial[13]
DMM	Dirichlet mixture model[13]	KDE	kernel descriptor [21]

A. Intermediate representations for scene recognition

A number of methods include mid-level representations using explicit classifiers. Vogel and Schiele [3] proposed a vocabulary with nine local concepts to model natural scenes. *Object bank* [4], [22] is a semantic representation that encodes the response at different spatial locations of a number of pretrained object classifiers. *Classemes* [23] are intermediate semantic representations based on a set of 2659 basis classes. These methods require training these intermediate classifiers explicitly (with the corresponding mid-level annotations) and often exploit large amounts of external training data (e.g., ImageNet, web images) to learn these mid-level classifiers.

Latent topic models are also a popular approach in which mid-level concepts are unknown and need to be discovered. They are often modeled using variants of *latent Dirichlet allocation* (LDA) [24], [25]. However, most LDA have been shown to capture irrelevant general regularities rather than the semantic regularities of interest, due to poor supervision [25]. Spatial context can be included to model the global layout and enforce local coherence in the topics [26], [7]. Recently, Li and Guo [27] proposed a patch-based latent framework which jointly learns the contextual representation and the classification model. Most latent topic models are generative, and usually do not scale well to large scale datasets. More recent variants exploit discriminative parts, which are unknown and discovered during learning [8], [9], [10], [28]. Alternatively, some variants [29], [30] learn the mid-level representations using dictionary learning.

B. Semantic multinomial

The *contextual multinomial* (CMN) [31], [13] uses the *semantic multinomial* (SMN) as intermediate representation for patches. Patch SMNs are learned via weak supervision using scene labels, common for all patches in each given image. To address the ambiguity (i.e., scene category co-occurrences) caused by this weak supervision, a second classifier (i.e., *contextual model* in [31], [13]) models the scene from SMNs and obtains the final classification. Note that this process has three advantages compared with other intermediate representations: a) no explicit mid-level vocabulary is required (not even a latent one), b) consequently, no expensive mid-level

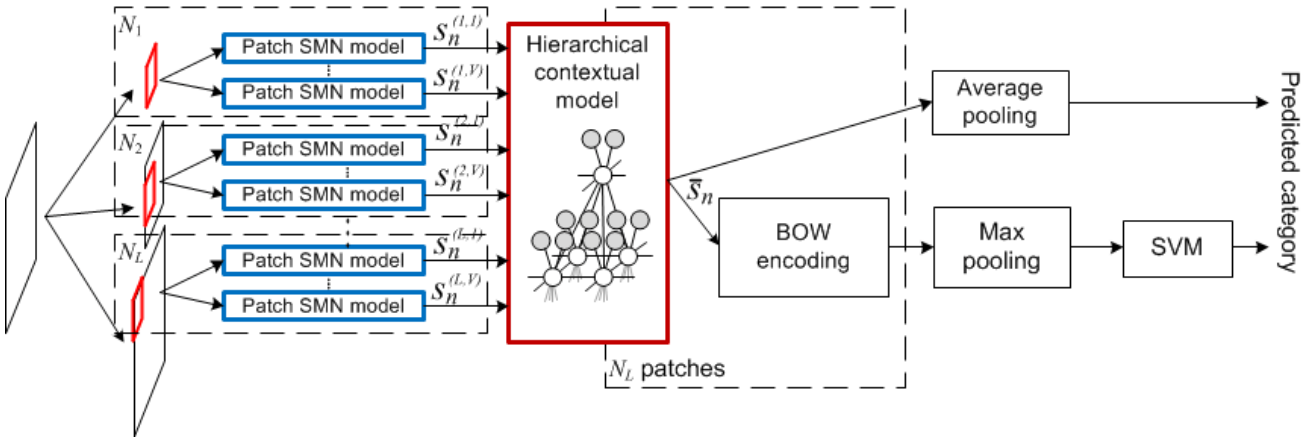


Fig. 2. Scene recognition framework, in test, we resize the input images to different L sizes of V kinds of visual feature.

TABLE II
COMPARISON OF SMN-BASED APPROACHES TO SCENE RECOGNITION. PROPOSED METHODS ARE EMPHASIZED IN BOLD.

Method	Patch features							Aggregation	Co-occurrence modeling		
	SMN	MF	S/D	Pre-train	WS	DSC	Speed		MS	Modeling tools	Classif
CMN[13]	GMM	No	S	-	Yes	No	Slow	Geometric	No	Global	DMM
(SPM)SM[11]	GMM	No	S	-	Yes	No	Slow	Voting	No	Global	SVM
CNF, KCNF, co-codes[19]	GMM	Yes	S	-	Yes	No	Slow	Average	No	Local+global	SVM
SFV[15]	IM-CNN[1]	-	D	IM	No	Yes	Fast	FV	Yes	-	SVM
Multi-feature spatial	GMM	Yes[2]	S	-	Yes	No	Slow	Average	No	MRF+local+global	SVM
	NN	Yes[2]	S	-	Yes	Yes	Medium	Average	No	MRF+local+global	SVM
	CNN	Yes[3]	D	IM/PL	Yes[4]	Yes	Fast	Average	No	MRF+local+global	SVM
Multi-scale multi-feature spatial	CNN	Yes[3]	D	IM/PL	Yes[4]	Yes	Fast	Average	Yes	MRF+local+global	SVM

S: shallow, **D**: deep, **MF**: multi-feature **MS**: multi-scale, **IM**: ImageNet, **PL**: Places, **WS**: weakly supervised, **DSC**: discriminative.
CMN: contextual multinomial, **(SPM)SM**: (spatial pyramid matching) semantic manifold, **MF**: multi-feature **(K)CNF**: (kernel) contextual noise filter, **SFV**: semantic Fisher vector.
[1] Patch-SMNs and image-SMNs do not lie on the same simplex, so there is no weak supervision nor scene category co-occurrences.
[2] Visual features: kernel descriptors (gradient, shape and color).
[3] Visual features: ImageNet-CNN and Places-CNN.
[4] Supervised pre-training on ImageNet or Places, and weakly supervised on the target data.

annotations are necessary, and c) still requires training models for the two stages, but in contrast to latent representations, these two stages can be trained independently (instead of jointly, which is more complex), making training much more scalable to large datasets.

The original CMN approach uses generative models, combining Gaussian mixture models (GMMs) and Dirichlet mixture models (DMMs). The *semantic manifold* (SM) [11] is a variant using the *negative geodesic distance* (NGD) kernel which is suitable for the geometry of the semantic manifold (i.e., a simplex), enabling discriminative classification with kernel SVMs instead of DMMs. Additionally, the SM can be extended with spatial pyramid matching (SPM) [20] for better classification (i.e., SPMSM). However, DMMs and kernel SVMs are still limited to relatively small datasets. For large scale classification, the (SPM)SM framework uses an approximate embedding of the NGD kernel [11], avoiding computing the kernel matrix, but at the cost of some accuracy. Recent extensions include unsupervised modeling[14] and better embeddings, such as the kernelized contextual noise filter (KCNF) [19]. However, these models still have limitations, which we address in this paper and describe in more detail in Section III-B.

C. Deep features

Deep convolutional neural networks (CNNs) [32] trained with large datasets [33], [18] are the current state-of-the-art feature representations, achieving impressive recognition performance in many visual recognition tasks, including object and scene classification [34], [18].

Recently, several weakly supervised frameworks [35], [36], [37], [38] have been proposed to detect and recognize objects. Oquab *et al.* [36] propose to fine tune pretrained CNNs with multiple regions, where a global max-pooling layer is used to select the regions for fine tuning. Durand *et al.* [35] extend this idea by selecting both "useless" (negative) and useful (positive) regions with a mixed (maximum and minimum) pooling layer. These weakly-supervised works focus on objects, requiring a pooling layer to select the most salient regions for object detection, while in this paper we focus on scene recognition, using all the patches for CNN fine tuning, and motivated from previous works using weak supervision on shallow features (e.g., GMM-SMNs) [13], [11].

Combining CNN features extracted at multiple scales can further improve the accuracy of scene classification [39], [15], [40], due to the wide range of objects appearing in scenes. Our framework also combines multiple scales and deep CNN features. Earlier, Gong *et al.* [39] extract CNN activations from patches, and encode them into multi-scale feature vectors

using VLAD. Wu *et al.* [40] extract deep features from a set of region proposals, which are then pooled into the scene representation. The semantic Fisher vector (SFV) approach [15], [41] uses Fisher vectors [42] to encode the output of the CNN. Note that the output of the softmax layer is a probability distribution, so it can be regarded as a SMN lying on the 1000-dimensional semantic space of the training categories (i.e., objects categories from ImageNet ILSVRC12). However, this intermediate semantic space is different from the final scene semantic space. In contrast, the semantic space in our case is common for both intermediate and scene representations, leading to scene category co-occurrences (see Table II).

III. The Semantic Manifold

A. Scene category co-occurrences

The *semantic multinomial* (SMN) descriptor $\mathbf{s} = (s_1, ..., s_M)^T$ [12] represents the probability $s_w = P(w|x)$ that a patch (or image) with visual feature x (e.g., SIFT [43], color histogram, kernel descriptors [21]) belongs to each scene category w, consisting of M scene categories in total. The term *semantic space* refers to the probability simplex where SMNs lie on. Since only image labels are available, patch models are learned using weak-supervision via image labels (see Fig. 3a). In particular, patches are modeled with GMMs $P_{\text{GMM}}(x|w)$ trained independently for each category w (i.e., only with patches from images in category w). Using the Bayes rule, each component of the SMN descriptor is obtained as the posterior probability $s_w = P_{GMM}(w|x) = P_{GMM}(x|w)\,P(w)\,/P(x)$. We will refer to SMNs obtained in this way as *GMM-SMNs*.

Patch SMNs in a given image are then aggregated into a single image SMN using their geometric average [13] or voting [11]. Weak-supervision during training creates a problem of ambiguity in the resulting image SMNs. For instance, patches containing pieces of *sky* can be found in images from many categories (e.g., *coast*, *mountain*, *highway*, *open country*). Since the visual content is very similar, all those patches may have visual features with similar distribution, but depending on the particular training image, the label will be different. Thus for an unknown test patch, the SMN descriptor will estimate certain probability in all those related categories. This can be seen as scene categories co-occurring in the SMN descriptor. Rasiwasia *et al.* [13] observed that some co-occurrence patterns are consistent across image SMNs in the same scene category, referring to them as *contextual co-occurrences*, masked by other undesirable co-occurrence patterns regarded as *contextual noise*. Thus, scenes can be modeled from these patterns, and hence the need for a second classifier (referred in [13] as *contextual classifier*).

B. Limitations

The original contextual multinomial and semantic manifold have several limitations that make them not competitive with the state-of-the-art in scene recognition:

(a) Category-specific patch GMMs are *redundant* and *not discriminative*. The reason is that they are trained independently per category, so each GMM ignores the other GMMs. Moreover, since all GMMs from all categories need to be evaluated to obtain patch representations, this also makes them *inefficient* for a large number of categories (i.e., the time to compute an image SMN grows as $O(NM)$).

(b) Image SMNs obtained with GMMs are very noisy, at both patch and image level, which leads to limited recognition performance since the image classifier cannot learn a suitable model from noisy data. While aggregating them into image SMNs can reduce the contextual noise, there is no specific method to filter contextual co-occurrences out from co-occurrence noise.

(c) Previous works CMN and (SPM)SM only exploit *global* contextual co-occurrences, i.e., those available in image SMNs, ignoring *local* ones at patch level. However, contextual co-occurrences are essentially local, and local modeling can significantly improve the recognition performance. While Song *et al.* [19] exploit co-occurrences in patches, still ignore the contextual relations between patches, which are key to remove co-occurrence noise.

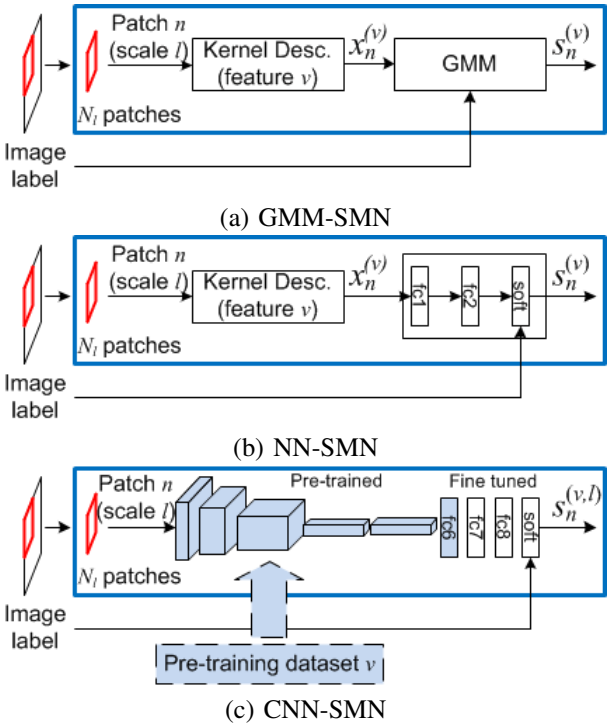


Fig. 3. Weakly-supervised learning of patch SMN models.

Addressing these limitations we include the following modifications (see framework in Fig. 2) in the semantic manifold framework:

- **Neural network-based patch** SMNs. Addressing limitation (a), we replace category-specific GMMs by a suitable multi-layer neural network model for patch SMNs. In contrast to GMMs, a single model is learned jointly for all categories in a *discriminative* way. Furthermore, representations in intermediate layers are shared, so they can scale much better to a larger number of categories. We study both *shallow* and *deep* architectures, the latter requiring pre-training with external data.

- **Multi-scale multi-feature contextual model.** We address limitations (b) and (c) with a hierarchical contextual model that exploits spatial, multi-feature and hierarchical relations between patches at different scales. In contrast to [14], we include explicit hierarchical relations between scales, and propose a message passing algorithm to better optimize the model.

IV. Discriminative semantic multinomials

A. Neural network based representations

A GMM can be considered as a multi-layer model with two levels of trainable parameters, i.e., the parameters of each Gaussian, and then the weights to combine them. As shown earlier, each model is learned independently for each category so they are not trained to discriminate between them. Besides, $P_{\text{GMM}}(x|w)$ is trained to fit the feature distribution of x for category w, which also includes those parts that are not discriminative. Thus, GMMs tend to require more parameters than a discriminative approach.

In order to obtain more discriminative SMN descriptors, we replace GMMs with a multi-layer neural network (NN) with the same depth (two layers). The neural network consists of two fully connected layer and one softmax layer (see Fig. 3b), where the output $P_{NN}(w|x)$ can be used directly as SMN (we refer to them as *NN-SMNs*). The network is still weakly-supervised via image labels. However, instead of having M independent GMMs, we have a single NN that jointly models all categories, thus being able to minimize a discriminative loss. By sharing intermediate layers, we can also learn more expressive models with comparable number of parameters. Table III shows a significant increase in the accuracy with NN-GMMs.

In addition to discriminability, NN-GMMs are more efficient and scalable to datasets with many categories, since only the last fully connected layer depends on the number of categories. The rest of the architecture can remain unchanged. In contrast, GMM-SMNs require training a new GMM for each additional category (i.e., the cost grows linearly with M, while NN-SMNs grow sublinearly). This results in significant speed-ups (e.g., around ten times faster for 67 categories, see Table III).

B. CNN based representations

We can further integrate the visual feature extraction stage as an additional layer(s) in a deeper model. In GMM-SMNs and NN-SMNs, the visual feature extraction stage is *handcrafted* (i.e., engineered, not trainable) while GMMs and the NN are trainable. Deep CNNs replace this first stage by several trainable convolutional layers. Similarly to NN-SMNs, we can use the output of a CNN architecture as $P_{CNN}(w|x)$, where now x are directly RGB pixels. The downsize is that these CNN models are significantly deeper with many more parameters. Since the training data is often limited, we train the SMN model in two steps. First, the CNN is pre-trained with a large dataset (e.g. ImageNet ILSRVC2012, Places). In practice we just reuse pretrained models. In order to adapt to the number of target scene categories, we replace the classifier implemented as the last fully connected layer by another fully connected one conveniently resized. We train this new layer (i.e. new classifier) and also fine tune the previous fully connected one (e.g. fc7 in VGGNet). As input we use the patches extracted at the corresponding scale, and as label we use the scene category of the corresponding image (i.e. weakly supervised training, as in GMM-SMNs and NN-SMNs). We refer to SMNs obtained with this method as *CNN-SMNs* (see Figure 3c).

Table III shows a significant gain compared with previous methods. The main reasons are the deeper model, a larger patch size and being trained end-to-end (at patch level). Note however that the CNN heavily relies on external large datasets used for pre-training. In addition, CNN-SMNs are significantly faster. Although the CNN model is more complex, it processes much fewer patches. Besides we implement patch extraction as convolutions (i.e., as a fully convolutional network), which is very efficient by reusing intermediate results in overlapping patches. Finally, visual feature extraction of kernel descriptors (KDES) is considerably slow since it is performed in the CPU while the other operations are performed in the GPU.

V. Context Model

A. Local category co-occurrences and contextual relations

A critical part in the SM framework is the contextual modeling of category co-occurrences to obtain robust classification. The original CMN and (SPM)SM approaches model category co-occurrences *after* aggregating patch SMNs into image SMNs. Thus they are limited to global co-occurrence patterns in image SMNs. While it can indeed address the ambiguity resulting from weak-supervision, image SMNs are still very noisy representations due to the fact that significant information is lost in this aggregation process. However, these methods ignore that category co-occurrences are essentially local and sparse [19], as shown in Figure 1c.

Furthermore, there are several types of contextual relations in both patch SMNs and image SMNs that we can exploit to further reduce the co-occurrence noise and emphasize consistent co-occurrences:

a) Class-specific patterns: Category co-occurrences appear with similar patterns in images from the same class. This is essentially the motivation of the contextual modeling in CMN [13] and (SPM)SM [11] approaches, which model scenes from these patterns in image SMNs. Song *et al.* [19] further exploit local co-occurrences and sparsity in an unsupervised way to filter co-occurrence noise in patch SMNs. They use dictionary coding in a bag-of-words fashion, which ignores explicit spatial relations between neighbors.

b) Local spatial relations: Neighboring patches are likely to depict parts of similar concepts (e.g., *sky*). Similarly, their patch SMNs are likely to have similar co-occurrence patterns.

c) Inter-feature relations: Since SMNs are semantic representations lying on the same semantic simplex (regardless of the input visual feature), different visual features generate complementary co-occurrence patterns after learning SMN models.

TABLE III
COMPARISON BETWEEN PATCH SMN MODELS FOR MIT INDOOR (67 SCENE CATEGORIES).

Method	Model	Image size	Patch size	Visual feature	#param	Total patches	Test time (s/image)			Acc. (%)
							Visual	SMN	Total	
GMM-SMN	67×GMM (512 components)	256×256	16×16	KDES (grad)	67K	30x30	0.51	1.12	1.63	34.7
NN-SMN	NN (67×512 hidden units)	256×256	16×16	KDES (grad)	34K	30x30	0.51	0.11	0.62	43.4
CNN-SMN	CNN (VGG, ImageNet)	448×448	224×224	Pixels	138M	8x8	0.07		0.07	71.0
	CNN (VGG, ImageNet)	640×640	224×224	Pixels	138M	14x14	0.15		0.15	73.4

Using our own implementation, NVidia TitanX GPU and Pentium i5 CPU. CNN pretrained on ImageNet ILSVRC2012. GMMs, NNs and CNNs are computed in the GPU.

d) Multi-scale relations: In a multi-scale setting, where images are resized to different scales, patches extracted from similar regions but different scales will still have certain similarity, so the corresponding SMNs and co-occurrence patterns will have too. We exploit them for the case of CNN-SMNs.

By properly exploiting jointly all these contextual relations between patch SMNs, consistent patterns can be emphasized while noisy accidental ones can be removed. In this paper we propose a context model that jointly addresses the four types of contextual relations.

In a first approach, we assume a single scale and a set V of complementary features (in our experiments $V = \{\text{gradient, shape, color}\}$ for GMM-SMN and NN-SMN, and $V = \{IM, PL\}$ for CNN-SMN, corresponding to the SMNs obtained with ImageNet-CNN and Places-CNN, respectively, adapted to the target scene categories as explained earlier). Each feature $v \in V$ generates a set of local visual descriptors $I^{(v)} = \{\mathbf{x}_1^{(v)}, \ldots, \mathbf{x}_N^{(v)}\}$, $\mathbf{x}_n^{(v)} \in \mathbf{X}^{(v)}$, and $I = \{I^{(1)}, \ldots, I^{(|V|)}\}$ represents all the features in the image. Now we assume that $P_v\left(\mathbf{x}_n^{(v)}|w\right)$ is the feature-specific patch model for feature v , learned independently in the same way as in the single feature case. Thus, we can define the feature-specific patch SMN of the patch n and the feature v as $\mathbf{s}_n^{(v)} = (s_{n1}^{(v)}, ..., s_{nM}^{(v)})^T$. Figure 4 shows an example with three feature-dependent patch SMNs. In this figure we can observe how certain regions are noisier than others in some features. We can also observe certain patterns across categories (category co-occurrences), across features (inter-feature relations) and between neighboring patches (spatial relations).

B. Global models

1) Single scale model: Since our objective is to keep consistent co-occurrences and remove accidental noise from patch SMNs, we formulate our contextual model as a denoising problem using a Markov Random Field (MRF).

Considering first a single feature and a 4-connectivity grid, the resulting model is shown in Figure 5b. The objective is to maximize the joint probability over the set of observed SMNs and denoised SMNs defined as $P\left(\bar{\mathbf{s}_1}, \ldots \bar{\mathbf{s}_N}, \mathbf{s}_1, \ldots \mathbf{s}_N\right) = \frac{1}{Z} exp\left(-E\left(\bar{\mathbf{s}_1}, \ldots \bar{\mathbf{s}_N}, \mathbf{s}_1, \ldots \mathbf{s}_N\right)\right)$, where Z is the partition function to normalize the probability. Thus, the problem is equivalent to minimizing the global energy of the network modeled as

$$E\left(\bar{\mathbf{s}_1}, \ldots \bar{\mathbf{s}_N}, \mathbf{s}_1, \ldots \mathbf{s}_N\right) = \sum_n g(\bar{\mathbf{s}_n}, \mathbf{s}_n) + \alpha \sum_{\{n,n'\}} g(\bar{\mathbf{s}_n}, \bar{\mathbf{s}_{n'}}) + \lambda H\left(\bar{\mathbf{s}_n}\right) \quad (1)$$

where $\bar{s_n}$ is the unknown denoised SMN of patch n (in contrast to the original s_n) and $\{n, n'\}$ represents pairs of connected patches. We model the energy as distance between SMNs. A suitable choice for the distance between points in simplices is the geodesic distance (GD) $g\left(\mathbf{s}, \mathbf{s}'\right)$[44]. We chose it over the KL divergence used in [13] because KL divergence is asymmetric, and in the semantic manifold framework GD has been proved effective [11]. Finally, we include a regularization term $H\left(\mathbf{s}\right) = -\sum_{w=1}^{M} s_w \log\left(s_w\right)$, which is the entropy of $\mathbf{s}$. This term is included to penalize too flat SMNs, which would lead to uninformative patches without co-occurrence patterns to model.

Considering now a multi-feature setting, all feature-dependent SMNs and the corresponding denoised SMNs lie on the same semantic space. Multi-feature combination can be easily achieved using some kind of pooling (e.g., (weighted) average) as in Figure 5a, but it would ignore spatial relations. In contrast, the previous MRF model can be easily extended to jointly consider multiple features using the model in Figure 5c. The corresponding energy is

$$E\left(\bar{\mathbf{s}_1}, \ldots \bar{\mathbf{s}_N}, \mathbf{s}_1^{(1)}, \ldots \mathbf{s}_N^{(1)}, \mathbf{s}_1^{(|V|)}, \ldots \mathbf{s}_N^{(|V|)}\right) = \rho \sum_n \sum_{v \in V} g\left(\bar{\mathbf{s}_n}, \mathbf{s}_n^{(v)}\right) + \alpha \sum_{\{n,n'\}} g\left(\bar{\mathbf{s}_n}, \bar{\mathbf{s}_{n'}}\right) + \lambda H\left(\bar{\mathbf{s}_n}\right) \quad (2)$$

2) Hierarchical model: Considering now a multiscale setting with L scales, we can further extend the MRF model to connect the patches at scale $l-1$ with the patches at scale $l = 1, \ldots, L$ in a hierarchical fashion. The size of patches increases from scale $l = 1$ to scale $l = L$. A global hierarchical model using 4-connectivity is illustrated in Fig. 6a. The resulting joint energy for an architecture with L scales is

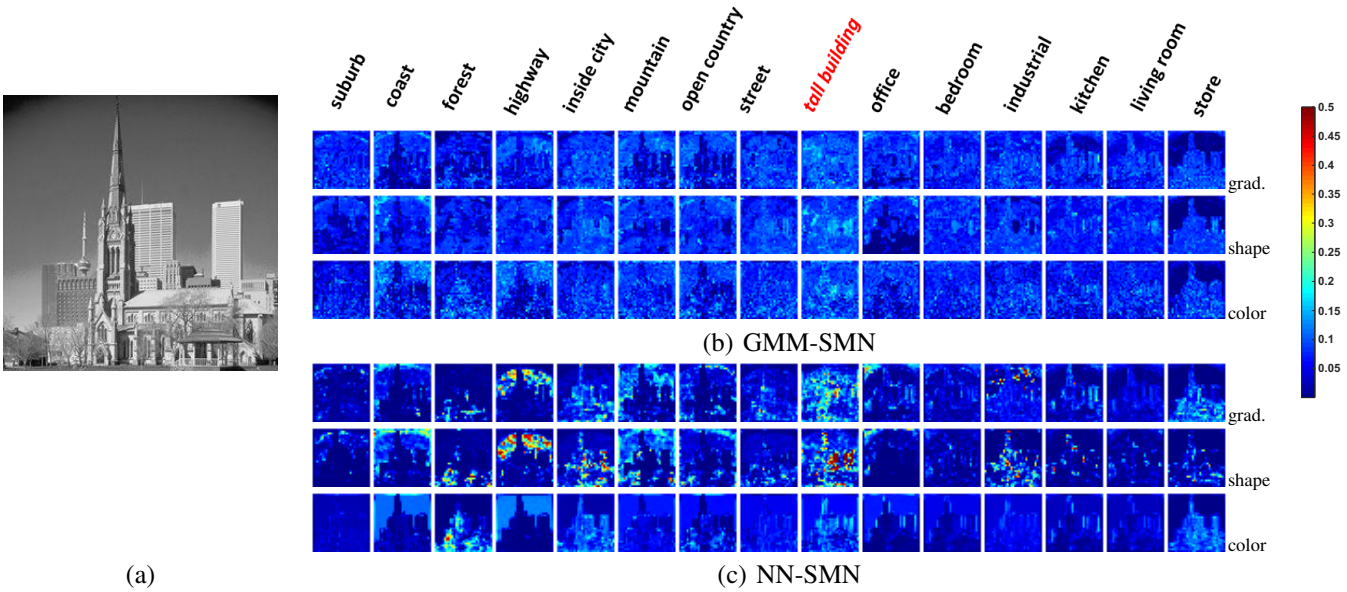


Fig. 4. Feature-specific patch SMNs as probability maps: (a) input image from the *15 scenes* dataset (category: *tallbuilding*), (b) GMM-SMNs, (c) NN-SMNs. Each row represents SMNs obtained from three different visual descriptors.

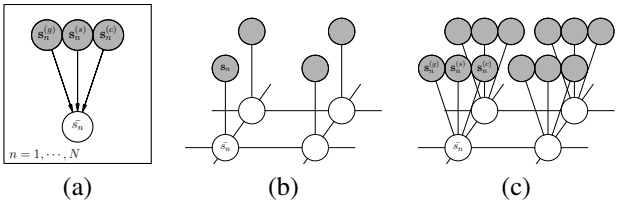


Fig. 5. Contextual models: (a) multi-feature combination, (b) 4-connected spatial grid model, and (c) multi-feature spatial grid model.

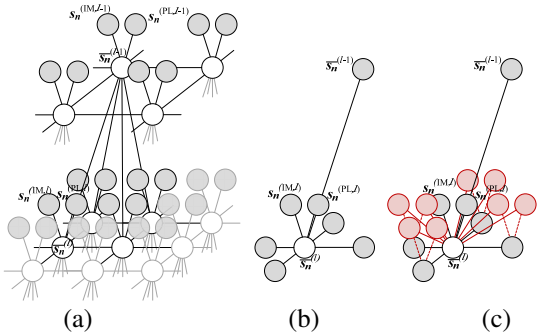


Fig. 6. Contextual models: (a) global, (b) local, and (c) extended local.

$$
\begin{aligned}
E(&\bar{\mathbf{s}_1}^{(1)}, \dots \bar{\mathbf{s}_{N_1}}^{(1)}, \bar{\mathbf{s}_1}^{(L)}, \dots \bar{\mathbf{s}_{N_L}}^{(L)}, \\
&\mathbf{s}_1^{(1,1)}, \dots \mathbf{s}_{N_1}^{(1,1)}, \mathbf{s}_1^{(L,|V|)}, \dots \mathbf{s}_{N_L}^{(L,|V|)}) = \\
&\rho \sum_{l=1}^{L} \sum_{n=1}^{N_l} \sum_{v \in V} g\left(\bar{\mathbf{s}_n}^{(l)}, \mathbf{s}_n^{(l,v)}\right) \\
&+ \alpha \sum_{l \in L} \sum_{\{n,n'\}} g\left(\bar{\mathbf{s}_n}^{(l)}, \bar{\mathbf{s}_{n'}}^{(l)}\right) \\
&+ \gamma \sum_{l=2}^{L} \sum_{(n,n^p)} g\left(\bar{\mathbf{s}_n}^{(l)}, \bar{\mathbf{s}_{n^p}}^{(l-1)}\right) + \lambda H\left(\bar{\mathbf{s}_n}\right) \quad (3)
\end{aligned}
$$

where N_l is the number of patches in scale l, and n^p in (n, n^p) represents the neighbor in previous scale.

To solve the minimization problem we can consider different alternatives commonly used in computer vision problems, such as image segmentation. However, we must emphasize the differences of our problem with image segmentation. In our case we are not interested in estimating the label of each patch, but in the probabilities in SMNs as scene features. Thus, algorithms designed to find the MAP labeling (e.g., graph cuts) are not easy to adapt to our problem. In the following subsections we describe different ways to address the optimization problem using different simplifications.

C. Local models

1) Hierarchical iterated conditional modes: The global hierarchical model can become very complex and difficult to optimize, particularly for the multi-scale scenario. In order to reduce the optimization complexity, we use a local approximation, inspired by the Iterated Conditional Modes (ICM) algorithm [16]. In this approximation, for a given patch $\bar{\mathbf{s}_n}^{(l)}$ the rest of the $\bar{\mathbf{s}_{n' \neq n}}^{(l)}$ are considered observed and fixed. Thus, the contextual model becomes local to $\bar{\mathbf{s}_n}^{(l)}$ and it is only necessary to consider a few connections. For example, the complex model of Fig. 6a is simplified to Fig. 6b. We use a hierarchical ICM algorithm that minimizes the global energy by scanning the patches and minimizing each local model one by one, updating the value of the corresponding $\bar{\mathbf{s}_n}^{(l)}$. For multiple scales, the scanning order is extended to include the multiple scales, following the top-down direction (see Algorithm 1). This algorithm can be seen as coordinate-wise gradient descent, and converges to a local optimum. The local energy for $\bar{\mathbf{s}_n}^{(l)}$ is computed as

$$E\left(\bar{\mathbf{s}_n}^{(l)};\phi_n^{(l)}\right)=\rho\sum_{v\in V}g(\bar{\mathbf{s}_n}^{(l)},\mathbf{s}_n^{(l,v)})+\alpha\sum_{(n,h),h\in B_n^{(l)}}g(\bar{\mathbf{s}_n}^{(l)},\bar{\mathbf{s}_h}^{(l)})+\gamma\sum_{(n,q),q\in Q_n^{(l-1)}}g\left(\bar{\mathbf{s}_n}^{(l)},\bar{\mathbf{s}_q}^{(l-1)}\right)+\lambda H\left(\bar{\mathbf{s}_n}\right)\quad(4)$$

$$\phi_n^{(l)}=\left\{\mathbf{s}_n^{(l,v)}|\forall v\in V\right\}\bigcup\left\{\mathbf{s}_h^{(l,v)}|\forall h\in B_n^{(l)},\forall v\in V\right\}\bigcup\left\{\bar{\mathbf{s}_q}^{(l-1)}|\forall q\in Q_n^{(l-1)}\right\}$$

where $B_n^{(l)}$ contains the neighbors in scale l, and $Q_n^{(l-1)}$ contains the related patches from the previous scale $l-1$. This local problem can be solved using gradient descent. The gradient corresponding to patch n at scale l is

$$\frac{\partial E'\left(\bar{\mathbf{s}_n}^{(l)};\phi_n^{(l)}\right)}{\partial s_n}=\rho\sum_{v\in V}\delta\left(\bar{\mathbf{s}_n}^{(l)},\mathbf{s}_n^{(l,v)}\right)+\alpha\sum_{(n,h),h\in B_n^{(l)}}\delta\left(\bar{\mathbf{s}_n}^{(l)},\bar{\mathbf{s}_h}^{(l)}\right)+\gamma\sum_{(n,q),q\in Q_n^{(l-1)}}\delta\left(\bar{\mathbf{s}_n}^{(l)},\bar{\mathbf{s}_q}^{(l-1)}\right)+\lambda H\left(\bar{\mathbf{s}_n}\right)\quad(5)$$

where

$$\delta\left(x,y\right)=\frac{\partial g\left(x,y\right)}{\partial x}=-\frac{\sqrt{y}}{2\sqrt{x}\sqrt{1-(\sqrt{x}\sqrt{y})^2}}$$

An advantage of this local model is that the complexity and computational cost are greatly reduced. We can easily extend this model to include relations with other neighboring SMNs (both observed and latent) without increasing significantly the complexity (see Fig. 6c). Note that these new neighboring relations are not part of the original global model of Fig. 6a. Including these extended relations in the global model and solving the optimization problem would be very difficult, since the extended connections destroy the factorization into pairwise factors, requiring higher order factors, and the corresponding energy terms in the formulation.

2) Scale-wise message passing algorithm: A different view of the hierarchical ICM is as neighboring patches sending update messages to the current patch, and then moving to the next one (see Fig. 7a). A message passing algorithm consists of update messages and a schedule for the updates. Algorithm 1 has the problem that each update may depend on both updated and not updated values. Here we study different message passing strategies as an alternative.

First we consider sending update messages directly in the global model. All nodes receive and send messages simultaneously in parallel, which are then updated at the same time as $\bar{\mathbf{s}_n}^{(l)}\leftarrow\bar{\mathbf{s}_n}^{(l)}+\triangle\bar{\mathbf{s}_n}^{(l)}$. The update is computed as

$$\triangle\bar{\mathbf{s}_n}^{(l)}=\rho\sum_{v\in V}\text{msg}\left(n^{(l,v)},n^{(l)}\right)+\alpha\sum_{(n,h),h\in B_n^{(l)}}\text{msg}\left(h^{(l)},n^{(l)}\right)+\gamma\sum_{(n,q),q\in Q_n^{(l-1)}}\text{msg}\left(q^{(l-1)},n^{(l)}\right)+\lambda H\left(\bar{\mathbf{s}_n}\right)\quad(6)$$

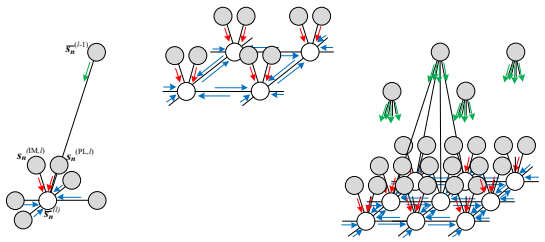
with $\text{msg}\left(h^{(l)},n^{(l)}\right)=\frac{\partial E'\left(\bar{\mathbf{s}_n}^{(l)};\phi_n^{(l)}\right)}{\partial s_n}$. Note that each message solves a local optimization problem as in previous section.

Since the information from top scales of CNNs is usually more reliable we devise a scale-wise message passing algorithm (Algorithm 2) that propagates the information from previous scales in a hierarchical fashion, rather than optimizing jointly the global model. The experiments will show that this strategy has better performance. The algorithm sends update

Algorithm 1 Hierarchical ICM

Input: Patch CNN-SMNs $\mathbf{s}_n^{(v,l)}$, $n=1,\ldots,N_l$, $l=1,\ldots,L$, $v=1,\ldots,|V|$
Output: Filtered patch CNN-SMNs $\bar{\mathbf{s}_n}^{(l)}n=1,\ldots,N_l$, $l=1,\ldots,L$

```
Initialize all s̄_n^(l) ← Σ_{v∈V} s_n^(v,l)/|V|
for l = 1 to L do
  for n = 1 to N_l do
    Calculate E(s̄_n^(l)) with Eq. 4
    // Local optimization for s̄_n^(l)
    for i = 1 to max_iter do
      energy_prev = E(s̄_n^(l))
      smn_prev = s̄_n^(l)
      Update s̄_n^(l) using Eq. 5
      Calculate E(s̄_n^(l)) with Eq. 4
      if E(s̄_n^(l)) ≥ energy_prev then
        s̄_n^(l) = smn_prev
        break
      end if
    end for
  end for
end for
return s̄_n^(l), n = 1,...,N_l, l = 1,...,L
```



(a) (b) (c)

Fig. 7. Optimization using message passing: (a) one step of hierarchical ICM, (b-c) two steps of the top-down scale-wise message passing algorithm for scale l and $l+1$.

messages within the nodes of a given single layer (including messages from the previous scale). In the next step, all the nodes in that scale are considered observed, and the next scale is processed in the same way. Fig. 7b and c represent two steps of this algorithm.

D. Embedding and pooling

After processing patch SMNs with the hierarchical context model, we aggregate them into image SMNs using average pooling, and the decision is simply the category with the maximum probability in the image SMN. Note that in this case the ambiguity due to weak supervision still remains. Alternatively, patch SMNs can be encoded and pooled prior to the contextual classifier [11], [19] (see Fig. 2). In particular we use the KCNF embedding [19] which exploits better local category co-occurrences.

Algorithm 2 Top-down message passing algorithm

Input: Patch CNN-SMNs $\mathbf{s}_n^{(v,l)}$, $n = 1, \ldots, N_l$, $l = 1, \ldots, L$, $v = 1, \ldots, |V|$
Output: Filtered patch CNN-SMNs $\bar{\mathbf{s}_n}^{(l)} n = 1, \ldots, N_l$, $l = 1, \ldots, L$

```
Initialize all s̄_n^(l) ← Σ_{v∈V} s_n^(v,l) / |V|
for l = 1 to L do
    Get the set D of all the edges connecting nodes in scale l and connecting nodes between l and l − 1
    for i = 1 to max_iter do
        for all (n', n) ∈ D do
            Calculate msg(n'^(l), n^(l))
        end for
        for n = 1 to N_l do
            Update all s̄_n^(l) using Eq. 6
        end for
    end for
end for
return s̄_n^(l), n = 1, ..., N_l, l = 1, ..., L
```

VI. Experiments

A. Experimental setup

1) Datasets: The proposed methods are evaluated on three small datasets. *15 scenes* [24], [20] contains 4485 images across 15 scene categories. *LabelMe*[1] consists of 8 outdoor scene categories, with a total of 2600 images. *UIUC-Sports* [5] consists of 1585 images labeled into 8 complex sport scene categories. Following the settings in previous works, we use 100, 100 and 70 images for training, respectively. We also evaluate the proposed methods on larger datasets, including MIT67 [8] and SUN397 [45]. MIT67 contains 15620 images of 67 indoor scene classes. SUN397 consists of 397 categories, with 108762 images in total. In the case of MIT67 Indoor and SUN397, the training/testing configurations are provided by the original authors. Finally, we also include an evaluation on the very large Places365-standard dataset [46] consisting of 365 scene categories, and 1,803,460 training images with the number of images per class varying from 3,068 to 5,000. We follow the public training/validation split for evaluation.

2) Shallow patch SMNs: We evaluate GMM-SMNs and NN-SMNs in the multi-feature setting with one scale and the proposed context models. As local visual descriptor we use three variants of kernel descriptors[21]: gradient, shape and color KDES. All local visual descriptors are extracted on a regular dense grid of 16×16 pixels (stride 8 pixels), resulting in 30×30 patch level local descriptors on a 256x256 image. For GMM-SMNs we train GMMs with 512 mixtures for each scene category. For NN-SMNs we use a network with two fully connected layers, including one hidden layer with 512 nodes. Note that this network has comparable amount of parameters to the model with 512 GMMs.

3) Deep patch SMNs: We use the VGG CNN architecture pre-trained either with ImageNet [17] or Places [18], replacing the size of the last fully convolutional layer fc8 to meet the number of categories. Then we fine tune the previous fully convolutional layer fc7 and train fc8 with the target datasets. Since the size of the patches is fixed in this architecture (224×224 pixels) , we extract features in four scales, obtained by resizing the input image to 224×224, 320×320, 448×448 and 640×640 pixels (scales 1, 2, 3 and 4, respectively). With these sizes we obtain 1×1, 4×4, 8×8 and 14×14 patches per scale, respectively.

B. Context models with shallow SMNs

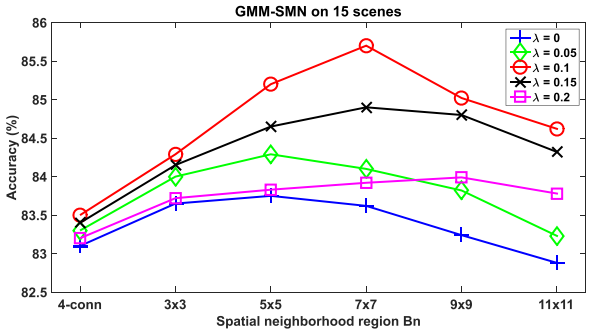
a) Baseline and proposed methods: We evaluate the proposed context models within the SM framework [11], but integrating KCNF encoding [19]. For GMM-SMNs and NN-SMNs we also include spatial pyramid matching [20] with four levels (1×1, 2×2, 3×3, 4×4).

Using the previous method as baseline, we consider four variations of the proposed context model:

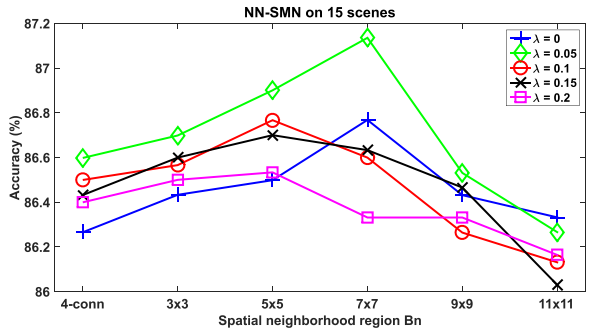
- *Multi-feature context* (MF): multiple features are combined in the semantic space with average pooling, corresponding to the model in Figure 5a.
- *Spatial context*: single feature exploiting neighboring spatial relations (see Figure 5b). Obtained by minimizing Eq. 1, when only one feature is used.
- *Multi-feature spatial context* (MFS): combines multiple-features of the target patch and neighboring spatial relations (i.e., see Figure 5c). Obtained by minimizing Eq. 2 in the multi-feature case.
- *Extended multi-feature spatial context* (EMFS): also includes multiple-features from additional patches in the neighborhood (Figure 6c with single scale).

1) Neighborhood size and entropy regularization: We evaluate the impact of the size of the spatial neighborhood which is critical in our context model. We use the *15 scenes* dataset and the EMFS model, fixing $\rho = 1/|V|$ and $\alpha = 1/|B_n^{(l)}|$. The results are illustrated in Figure 8.

We evaluate different neighborhoods, including the 4-connectivity spatial neighborhood, and other dense neighborhoods of size L×L patches (3x3 corresponds to 8 neighbors). We can observe that larger neighborhoods can effectively reinforce consistent patterns and filter accidental ones. However,



(a)



(b)

Fig. 8. Region size and sparse parameter evaluation

too large neighborhoods cannot capture properly local co-occurrence patterns. From our experiments, a good trade-off is 7×7 patches.

Entropy regularization is also important to capture category co-occurrence patterns properly. We evaluate λ in a range from 0 to 0.2, with a step of 0.05. Figure 9 shows that without entropy regularization ($\lambda = 0$) the performance is lower. Note that NN-SMNs require lower penalty than GMM-SMNs. We obtained the best performance for $L = 7$ and $\lambda = 0.1/0.05$ for GMM-SMN and NN-SMN, respectively, so for the rest of the experiments we use this configuration.

Figure 9 illustrates how the proposed method is able to effectively combine the three feature-specific patch SMNs from Figure 4 into smoother multi-feature patch SMNs. The regularization term prevents from excessive smoothing that can wash out the true class-specific co-occurrence patterns that we want to preserve.

2) Context models: We compare the different variations of the proposed method on the three small datasets to show how different types of context models improve the accuracy. Table IV shows that the classification accuracy increases consistently when we include additional contextual relations in the context model. Combining multiple features helps with a gain around 1.1-2.5%/1.3-1.4% (GMM-SMN/NN-SMN) over the best single feature. Using spatial relations varies from no gain to modest gains around 1%/3.3%. However, combining both can increase an additional 0.5-1%/0.7-0.8% over only multi-feature context. The extended multi-feature spatial context contributes with an additional 0.4-2.2%/1.1-2.6% gain by incorporating multiple features from the neighboring patches. The total gain with the extended context model over the baseline is around 2.6-5.7%/3.0-4.5%. Note also that NN-SMNs typically obtain slightly better performance compared with GMM-SMNs, and both consistently benefit from contextual modeling.

TABLE IV
ACCURACY (%) OF GMM-SMN/NN-SMN FOR DIFFERENT CONTEXT MODELS. * INDICATES IMPLEMENTED BY US INSTEAD OF REPORTED.

Method (feature)	15 scenes	LabelMe	Sports
No context model (Baseline)			
Gradient)	78.9/82.1	86.5/86.5	83.9/84.3
Shape)	80.0/82.5	85.0/85.4	84.3/85.1
Color	75.4/72.3	72.4/70.0	72.8/71.9
Spatial context (7×7 patches)			
Gradient	81.0/82.6	86.7/86.9	83.7/84.6
Shape	81.4/82.7	84.9/85.7	83.9/85.5
Color	76.6/75.5	72.9/74.2	73.1/75.2
Multiple feature context			
Multi-feature	82.5/83.8	88.3/87.9	85.4/86.5
Joint multi-feature spatial context (7×7 patches, $\lambda = 0.1/0.05$)			
MFS	83.5/84.5	88.9/88.8	85.9/87.3
Extended MFS	**85.7/87.1**	**89.3/89.9**	**86.9/88.5**
Related works			
CMN[13]	77.2	-	-
LDA[25]	76.6	-	-
SPMSM[11]	82.5	87.5	83.0
Kernel descriptor*	82.2	87.3	85.2
Object-Bank[27]	85.7	89.8	83.9
KCNF [19]	85.2	89.8	87.8
Object-to-Class kernels[22]	88.8	-	86.0

3) Comparison with related works: We compare our method with other works using mid-level semantic representations, such as latent topics [25] *object bank* [4], [27], [22] and *classemes* [47], [23]. Most of these approaches cannot be used in large scale datasets, so we separate comparisons for small datasets and larger datasets.

a) Small datasets: Table IV compares the results reported by the authors in their corresponding references. Although a completely fair comparison with reported results is not possible, due to different implementations, features and other parameters, our framework at least seems to be competitive in the three evaluated datasets. Comparing with previous methods based on SMNs and co-occurrence modeling, such as CMN, SPMSM and KCNF, is of particular interest. The proposed method, which also exploits multiple features and richer contextual relations, achieves better performance than those methods. We also compare with non-semantic representations by directly modeling categories from the same low-level kernel descriptors (concatenated to combine them), with and a SVM and spatial pyramid. We observe that our method also achieves better results.

b) Large datasets: We evaluate the proposed methods on the larger MIT67 and SUN397 datasets. The results are shown in Tables V and VI, respectively. NN-SMNs achieve better performance than GMM-SMNs, especially for MIT67. The gains due to richer context models are much higher than in smaller datasets, with significant gains of 11%/9.5% and 15%/9.1% (GMM-SMNs/NN-SMNs) over the best single feature baseline, respectively. This suggests that contextual relations become much more important important as the number of scene categories increases, resulting in much noisier and sparser co-occurrence patterns. Exploiting the context to

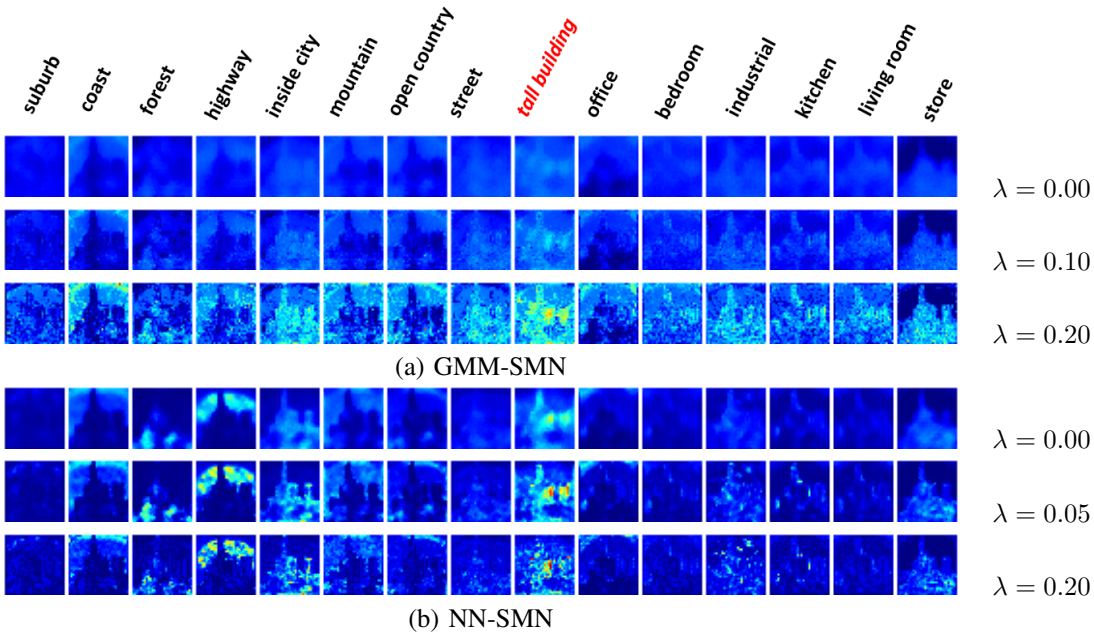


Fig. 9. Output patch SMNs of the image in Fig 1a (category: tallbuilding) after the context model and the effect of entropy regularization: (a) GMM-SMNs, and (b) NN-SMNs. The spatial neighborhood is 7×7 patches.

TABLE V
COMPARISON ON MIT67 DATASET.

MIT67	Method	Accuracy (%) GMM-SMNs	Accuracy (%) NN-SMNs
Proposed	Baseline (gradient)	34.7	43.8
	Baseline (shape)	36.9	44.7
	Baseline (color)	26.8	29.4
	Proposed (MF)	42.4	47.2
	Proposed (MFS)	44.7	50.5
	Proposed (EMFS)	**48.2**	**54.3**
Other semantic representations	ObjectBank[27]	37.6	
	Object-to-Class kernels[22]	39.6	
	Deformable Part Models[48]	43.1	
	SPMSM[11]	44.0	
	KCNF [19]	48.1	
	Discriminative parts[9]	64.0	
Bag-of-words representations	Sparse Spatial Coding[49]	44.4	
	Geometric Phrase Pooling[50]	46.4	
	Linear Distance Coding[51]	46.7	
	IFV[52]	60.8	

TABLE VI
COMPARISON ON SUN397 DATASET.

SUN397	Method	Accuracy (%) GMM-SMNs	Accuracy (%) NN-SMNs
Proposed	Baseline (gradient)	25.4	30.9
	Baseline (shape)	23.2	32.6
	Baseline (color)	18.2	21.4
	Proposed (MF)	30.4	37.5
	Proposed (MFS)	34.9	39.7
	Proposed (EMFS)	**40.7**	**41.7**
Other semantic representations	SPMSM[11]	28.2	
	Meta-classes[23]	36.8	
	KCNF [19]	40.8	
Others	SUN (HOG)[45]	27.2	
	SUN(MKL)[45]	38.0	
	IFV[52]	47.2	

emphasize representative category co-occurrence patterns can greatly help to improve the recognition performance. Other mid-level semantic representations, such as *object bank* and *meta-classes* exploit larger amounts of external data (e.g.,. ImageNet, web images) to model the mid-level classifiers. The proposed method outperforms them without resorting to external data, but still falls short compared with discriminative parts [9], which is particularly effective for indoor scenes where certain objects can be very discriminative. However, this method cannot scale to larger datasets such as SUN397.

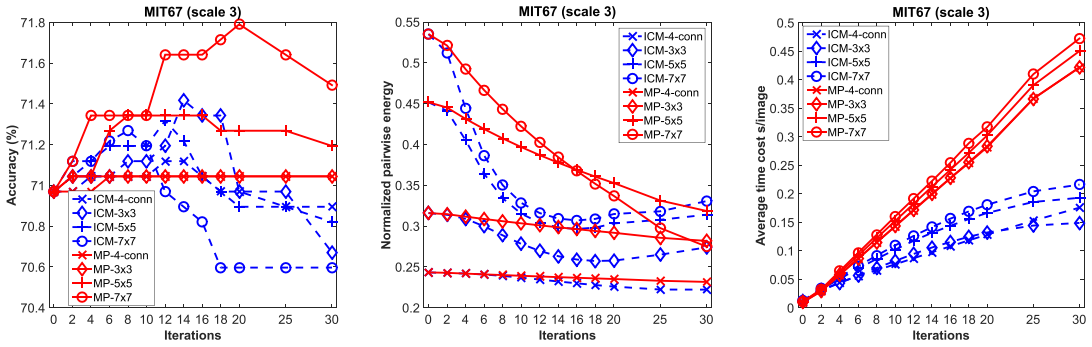
We also include other approaches based on lower level representations, such as bag-of-words coding [51], [49], [50] and the Fisher vector [52]. The latter achieves better accuracy, but at the cost of a much higher dimensional feature resulting from a much denser grid for sampling local features [52].

C. Context models with deep SMNs and multiple scales

1) Patches vs full images : We use the CNN-SMNs as described in Section IV-B, extracting two complementary features that depend on the pre-training dataset (i.e., either ImageNet-CNN or Places-CNN). In addition we consider multiple scales, which are determined by the size the input image is resized (for a fixed patch size of 224×224 pixels).

When adapting the CNN to a particular target scene dataset, this adaptation can be performed using full size images (resized to the patch size, i.e., 224×224 pixels) or using patches extracted at the particular scale. As Table VII shows, the latter is a better approach, since patches used for adaptation and during test have similar scale distributions.

2) Single scale: We first compare the hierarchical ICM and the message passing (MP) algorithms in a single scale setting. We compare the accuracy and the total energy for different spatial neighborhoods. Since the total energy depends on the number of edges, and they depend on the size of the neighborhood, it is difficult to compare neighborhoods with different



a b c

Fig. 10. Comparison between ICM and MP on MIT Indoor 67 on scale 3 (448x448), (a)accuracy, (b) normalized pairwise energy,(c) time cost

TABLE VII
ACCURACY (%) OF DIFFERENT ADAPTATIONS ON MIT INDOOR 67

scale	Images (fixed)		Patches	
	IM	PL	IM	PL
1	69.9	80.5	69.9	80.5
2	69.8	72.1	70.8	79.0
3	68.6	71.0	71.0	72.1
4	64.6	51.0	73.4	65.7

size. For better comparison, we normalized the energy and set $\rho = 1/(3|V|)$, $\alpha = 1/\left(3|B_n^{(l)}|\right)$, $\gamma = 1/\left(3|Q_n^{(l-1)}|\right)$ and $\lambda = 0$ in Eq. 4 and 5, which we found work well in practice. Fig. 10b shows how the energy of ICM decreases quickly to the minimum value in around 16 iterations. However, it increases with more iterations probably due to the asynchronous updating scheme, also causing the accuracy to decrease. In contrast, MP passes messages synchronously and then updates the values of each node simultaneously. As a result, the energy decreases more slowly but consistently (although the absolute value of the of the energy is slightly higher) and the accuracy increases slightly. However, a drawback is that is slower than ICM.

3) Multiple scales and message passing: In the next experiment we evaluate three variants of the proposed multi-scale MP algorithm on MIT Indoor 67 with just one CNN or combining two (both ImageNet-CNN and Places-CNN). The *integrated* variant optimizes all the nodes at the same time, and then combines the scales. The *top-down* and *bottom-up* variants are scale-wise, and progressively update a given scale based on the previous scale. In general, the *top-down* strategy performs better than the others, since the top scale (more global) obtains the best single-scale performance, so using it as initial step leads to a better solution.

The results of the same experiment for SUN397 are shown in Table IX. The proposed architecture combining ImageNet-CNN, Places-CNN at three scales achieves a remarkable 69.3% of accuracy, comparable to human performance, as reported in [53]. In this case including scale 4 decreases the performance, so we do not include it in the next experiment.

4) Encoding methods and other works: In the previous experiments there is no supervised contextual classifier (e.g., SVM) nor any particular encoding. The scene prediction is obtained basically pooling patch CNN-SMNs. Now we also consider the full SM framework, which includes encoding and SVM classification (see Figure 2). We selected the architectures with best performance from previous experiments (3 scales for MIT67 and 4 scales for SUN397, both with ImageNet-CNN and Places-CNN) and encode the CNN-SMNs using various encodings (SM [11], FV [15], EMK [54], LLC [55]). For EMK and LLC we use dictionaries of 1000 words, and for FV we use 50 GMMs and then reduce the dimensions to 4096 using PCA, following [15]. The results are shown in Table X. The gain using encoding+SVM is more significant for MIT67 than for SUN397, and for single scale than for multiple scales. In particular for SUN397, a marginal 0.1% gain is achieved over the best performance.

We also compare with other works in Table X, some using AlexNet and some using VGG architectures. In the next section we evaluate our approach on the recent dataset Places365[46]. Thus, we can also use CNNs pretrained on this dataset in our framework, and we report some results using an extended framework with in addition to ImageNet-CNNs and Places-CNNs includes Places365-CNNs. This setting obtains state-of-the-art performance 72.6% for SUN397.

D. Evaluation on Places365

Evaluation on Places365 is difficult due to the size of the dataset. In this case, we use the original crops for adaptation instead of patches and 3 scales (the amount of data resulting for smaller patches is too large for training). However, even with these settings the results of our framework with multi-scale multi-CNN context modeling obtains 57.1% top 1 accuracy, outperforming the best in baseline by 2.2%. We can also compare with a simple average pooling across scales and CNNs. and where our model still has a gain of 1.6%.

In general, evaluation on Places or Places365 is not reported in the vast majority of papers about scene recognition, and even most recent works typically use off-the-shelf CNNs trained on Places or Places365 but do not evaluate on those datasets. For Places365 we are only aware of the result of Zhou *et al*[46], which we improve by 1.9%. Note that their setting would be closer to our scale 1* (256×256 pixels), but with some differences: [46] averages 10 crops (4 corners+central+mirror), while we use only four (2×2 patches).

TABLE VIII
COMPARISON BETWEEN INTEGRATED AND SCALE-WISE MP MODELS ON MIT67 IN ACCURACY (%)

#scales	Scale	Integrated			Top-down			Scale	Bottom-up		
		IM	PL	IM+PL	IM	PL	IM+PL		IM	PL	IM+PL
1	1	69.9	80.5	**82.6**	69.9	80.5	**82.6**	4	73.4	65.7	76.4
2	1,2	73.1	82.6	83.2	73.6	82.1	**83.7**	4,3	73.8	74.4	78.4
3	1,2,3	75.7	82.3	83.3	77.5	82.1	**84.1**	4,3,2	73.8	80.4	81.9
4	1,2,3,4	75.8	82.1	83.6	78.7	81.6	**84.2**	4,3,2,1	74.6	81.3	82.9

TABLE IX
ACCURACY (%) OF MODELING JOINT CONTEXTS ON SUN397

Scale	Adaptation				Scale	Contextual modeling		
	With images		With patches			Spatial		IM+PL
	IM	PL	IM	PL		IM	PL	
1	55.9	65.8	55.9	65.8	1	55.9	65.8	67.1
2	55.8	62.3	55.8	62.5	1,2	57.9	66.0	68.4
3	49.9	46.5	55.2	56.2	1,2,3	60.8	66.8	**69.3**
4	36.5	20.3	54.5	44.0	12,3,4	60.1	66.8	68.6

VII. CONCLUSIONS

Although recently relegated in favor of deep learning methods, intermediate representations have played an important role in automatic scene recognition. The semantic manifold framework addresses the problem of modeling scene categories from visual features with a combination of weak supervision and pooling, that avoids mid-level annotations while inference can be easily modeled in two independent steps (in contrast to most methods that learn latent representations). This framework suffers from the specific problem of scene category co-occurrences, thus requiring specific solutions.

In this paper we revisit the semantic manifold approach and tackle several of the limitations not addressed in previous works [13], [11], [19]. We identify the original patch SMN models based on GMMs (i.e., GMM-SMNs) as an important bottleneck in terms efficiency and accuracy, resulting from the training stage that learns patch SMN models independently for each category. We show that replacing them by NN-SMNs, based on neural networks and learned jointly for all the categories, produce much faster and more discriminative SMNs.

Modeling category co-occurrences properly is the other critical stage. Previous methods ignore local contextual relations, which are very helpful for this purpose. SMN representations in the semantic manifold have the unique characteristic that patches and images are represented in the same semantic space, independently of the visual feature used as input. Exploiting this property, we combine multiple features and scales, and integrate spatial, multi-feature and even multi-scale relations between neighboring patch SMNs into a joint context model, showing that in this way we can discover consistent co-occurrence patterns and filter out noisy ones, making things easier for the classifier, which can focus on modeling scenes in terms of these cleaner patterns. In particular we use a multi-feature multi-scale Markov random field formulation, with a specific entropy regularizer. Although still far from CNNs and some methods, using the proposed NN-SMNs and an extended context model, our framework can significantly improve the recognition performance of the previous semantic manifold approach and its variants.

We further recast convolutional networks as sophisticated SMNs, implemented as weakly supervised adaptation of a pre-trained network, and integrate them as semantic features in the proposed framework. This hybrid approach achieves state-of-the-art scene recognition accuracy (even without the contextual classifier).

REFERENCES

[1] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001.

[2] J. Wu and J. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, Aug 2011.

[3] J. Vogel and B. Schiele, "Semantic modeling of natural scenes for content-based image retrieval," *Int. J. Comput. Vision*, vol. 72, no. 2, pp. 133–157, Apr. 2007.

[4] L. Li, H.Su, E. Xing, and L. Fei-Fei, "Object bank: A high-level image representation for scene classification and semantic feature sparsification," in *NIPS*, 2010.

[5] L. F.-F. L.J. Li, "What, where and who? classifying events by scene and object recognition," in *ICCV*, 2007.

[6] C. Wang, D. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *CVPR*, 2009, pp. 1903–1910.

[7] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context aware topic model for scene recognition," in *CVPR*, 2012.

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

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

[10] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in *CVPR*, 2013.

[11] R. Kwitt, N. Vasconcelos, and N. Rasiwasia, "Scene recognition on the semantic manifold," in *ECCV*, 2012.

[12] N. Rasiwasia and N. Vasconcelos, "Bridging the gap: Query by semantic example," *IEEE Trans. on Multimedia*, vol. 9, no. 5, pp. 923–938, 2007.

[13] ——, "Holistic context models for visual recognition," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 34, no. 5, pp. 902–917, 2012.

[14] X. Song, S. Jiang, and L. Herranz, "Joint multi-feature spatial context for scene recognition on the semantic manifold," in *CVPR*, June 2015.

[15] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos, "Scene classification with semantic fisher vectors," in *CVPR*, 2015.

[16] J. Besag, "On the statistical analysis of dirty pictures," *Journal of the Royal Statistical Society. Series B*, vol. 48, no. 3, pp. 259–302, 1986.

[17] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 1–42, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11263-015-0816-y

[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *NIPS*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Weinberger, Eds., 2014, pp. 487–495.

[19] X. Song, S. Jiang, L. Herranz, Y. Kong, and K. Zheng, "Category co-occurrence modeling for large scale scene recognition," *Pattern Recognition*, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316000406

[20] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

TABLE X
COMPARISON TO THE STATE-OF-THE-ART. *INDICATES OUR IMPLEMENTATION

Method	Encoding	Classifier	Dimension	MIT67/SUN397 Accuracy(%) Single scale	Multi-scale
Proposed (ICM)	none	none	67/397	76.4/61.5	83.1/68.1
Proposed (MP)	none	none	67/397	77.8/61.9	84.2/69.3
	SM[11]	SVM	67/397	81.0/62.9	**86.9**/69.4
	FV[15]	SVM	4096	81.8/62.9	86.1/69.4
	EMK[54]	SVM	1000	78.1/60.8	84.4/68.7
	LLC[55]	SVM	1000	76.8/60.0	83.9/68.0
Proposed (MP+PL365)	none	none	67/397	81.6/68.8	85.2/72.5
	SM [11]	SVM	67/397	82.1/69.1	86.5/**72.6**
Handcrafted	IFV [52]	SVM	128000	60.8/47.2	
CNN (AlexNet)	Hybrid-CNN (Fc7) [18]	SVM	4096	70.8/53.9	
	CNN (Fc7)+VLAD [39]	SVM	12288	68.9/52.0	
	CNN+SINN [56]	softmax	4096	-/57.6	
	MetaObject [40]	SVM	4096	78.9/58.1	
	Semantic FV [15]	SVM	20480	79.0/61.7	
CNN (VGG)	ImageNet-Deep19 [57]	softmax	-	70.8/51.9	
	DAG-CNN [58]	SVM	9216	77.5/56.2	
	WELDON [35]	softmax	67/-	78.0/-	
	Multiscale CNN-SMN pooling*	none	67/397	84.5/67.8	
	Semantic FV [15]*	SVM	20480	85.4/68.3	
	MFA-FS+Places(VGG) [59]	SVM	19096	87.2/71.1	
	VGG Places365	none	67/397	79.8/70.2	
Human-level scene recognition [45]		-	-	-/68.5	

Proposed (MP) 3-CNNs indicates the combination of ImageNet-VGG, Places205-VGG and Places365-VGG

TABLE XI
ACCURACY (%) OF VALIDATION DATASET ON PLACES365

Baseline				**Proposed**		
Scale	Single CNN			Scale	Multiple CNNs	
	IM-VGG	PL365-VGG	PL365-ResNet		Pooling	Context
1	47.8	53.1	52.6	1	53.8	54.3
1*	48.6	54.9	54.5	1,1*	55.5	56.8
2	49.4	54.3	54.1	1,1*,2	55.4	**57.1**
Related work						
PL365-VGG [46]					55.2	
PL365-ResNet [46]					54.7	

1*Indicates 256×256 pixels of input, resulting 2×2 patches.

[21] L. Bo, X. Ren, and D. Fox, "Kernel descriptors for visual recognition," in *NIPS*, 2010.

[22] L. Zhang, X. Zhen, and L. Shao, "Learning object-to-class kernels for scene classification," *IEEE Trans. on Image Process.*, vol. 23, no. 8, pp. 3241–3253, Aug 2014.

[23] A. Bergamo and L. Torresani, "Classemes and other classifier-based features for efficient object categorization," in *IEEE Trans. on Pattern Anal. and Mach. Intell.*, 2014.

[24] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *CVPR*, 2005.

[25] N. Rasiwasia and N. Vasconcelos, "Latent dirichlet allocation models for image classification," *IEEE Trans. on Pattern Anal. and Mach. Intell.*, vol. 35, no. 11, pp. 2665–2679, 2013.

[26] X. Wang and E. Grimson, "Spatial latent dirichlet allocation," in *NIPS*, 2007.

[27] L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition," *Int. J. Comput. Vision*, vol. 107, no. 1, pp. 20–39, 2014.

[28] X. Bai, C. Yao, and W. Liu, "Strokelets: A learned multi-scale mid-level representation for scene text recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 6, pp. 2789–2802, June 2016.

[29] X. Wang, B. Wang, X. Bai, W. Liu, and Z. Tu, "Max-margin multiple-instance dictionary learning," 2013, pp. 846–854.

[30] G. S. Xie, X. Y. Zhang, S. Yan, and C. L. Liu, "Hybrid cnn and dictionary-based models for scene recognition and domain adaptation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. PP, no. 99, pp. 1–1, 2015.

[31] N. Rasiwasia and N. Vasconcelos, "Holistic context modeling using semantic co-occurrences," in *CVPR*, 2009, pp. 1889–1895.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1106–1114.

[33] J. Deng, A. Berg, and L. Fei-Fei, "Large graph construction for scalable semi-supervised learning," in *ICML*, 2011.

[34] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in *ICML*, 2014.

[35] T. Durand, N. Thome, and M. Cord, "Weldon: Weakly supervised learning of deep convolutional neural networks," in *CVPR*, June 2016.

[36] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, "Is object localization for free? - weakly-supervised learning with convolutional neural networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[37] X. Li, L. Zhao, L. Wei, M. H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919–3930, Aug 2016.

[38] H. Bilen and A. Vedaldi, "Weakly supervised deep detection networks," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[39] Y. Gong, L. Wang, R. Guo, and S. Lazebnik., "Multi-scale orderless pooling of deep convolutional activation features," in *ECCV*, 2014.

[40] R. Wu, B. Wang, W. Wang, and Y. Yu, "Harvesting discriminative meta objects with deep cnn features for scene classification," in *ICCV*, 2015.

[41] Q. Wang, P. Li, W. Zuo, and L. Zhang, "Raid-g: Robust estimation of approximate infinite dimensional gaussian with application to material recognition," in *CVPR*, June 2016.

[42] F. Perronnin, J. Sanchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.

[43] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, pp. 91–110, 2004.

[44] D. Zhang, X. Chen, and W. S. Lee, "Text classification with kernels on the multinomial manifold," in *RDIR*, 2005, pp. 266–273.

[45] J. Xiao, J. Hayes, K. Ehringer, A. Olivia, and A. Torralba, "SUN database: Largescale scene recognition from abbey to zoo," in *CVPR*, 2010.

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

[47] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *ECCV*, 2010.

[48] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *ICCV*, 2011.

[49] G. Leivas Oliveira, E. Nascimento, A. Wilson Vieira, and M. Montenegro Campos, "Sparse spatial coding: A novel approach to visual recognition," *IEEE Trans. on Image Process.*, vol. 23, no. 6, pp. 2719–2731, June 2014.
[50] L. Xie, Q. Tian, M. Wang, and B. Zhang, "Spatial pooling of heterogeneous features for image classification," *IEEE Trans. on Image Process.*, vol. 23, no. 5, pp. 1994–2008, May 2014.
[51] Z. Wang, J. Feng, S. Yan, and H. Xi, "Linear distance coding for image classification," *IEEE Trans. on Image Process.*, vol. 22, no. 2, pp. 537–548, Feb 2013.
[52] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek, "Image classification with the fisher vector: Theory and practice," *Int. J. Comput. Vision*, vol. 105, no. 3, pp. 222–245, 2013.
[53] J. Xiao, K. Ehinger, J. Hays, A. Torralba, and A. Oliva, "Sun database: Exploring a large collection of scene categories," *International Journal of Computer Vision*, pp. 1–20, 2014.
[54] L. Bo and C. Sminchisescu, "Efficient match kernel between sets of features for visual recognition," in *NIPS*, 2009.
[55] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *CVPR*, 2010.
[56] H. Hu, G.-T. Zhou, Z. Deng, Z. Liao, and G. Mori, "Learning structured inference neural networks with label relations," in *CVPR*, June 2016.
[57] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[58] S. Yang and D. Ramanan, "Multi-scale recognition with DAG-CNNs," in *ICCV*, 2015.
[59] M. D. Dixit and N. Vasconcelos, "Object based scene representations using fisher scores of local subspace projections," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 2811–2819.