Exploring Confusing Scene Classes for the Places Dataset: Insights and Solutions

Chen Chen Shangwen Li Xiang Fu Yuzhuo Ren Yueru Chen C.-C. Jay Kuo
Department of Electrical Engineering
University of Southern California, Los Angeles, CA 90089, U.S.A.
{chen80, shangwel, xiangfu, yuzhuore, yueruche}@usc.edu, cckuo@sipi.usc.edu

***Abstract*—Scene classification is more challenging than object classification due to higher ambiguity in scene labels. In this work, we propose to use the filter weights at the last stage of a CNN model trained by the Places dataset, which is also known as the "scene anchor vector (SAV)", to explain the source of confusions. An SAV points to a cluster of images. If two anchor vectors have a smaller angle, we see overlapping image clusters, leading to a set of confusing classes. To overcome it, we propose to merge images associated with confusing anchor vectors into a confusion set and split the set in an unsupervised fashion to create multiple subsets. It is called the "automatic subset clustering (ASC)" process. Each of these subsets contains scene images of strong visual similarity. After the ASC process, we train a random forest (RF) classifier for each confusion subset to allow better scene classification. The ASC/RF scheme can be added on top of any existing scene-classification CNN as a post-processing module with little extra training effort. It is shown by extensive experimental results that, for a given baseline CNN, the ASC/RF scheme can offer a significant performance gain.**

I. Introduction

The Convolutional Neural Network (CNN) has been wildly applied to large-scale visual recognition problems such as object and scene image classification in recent years. Its popularity grows rapidly with the emergence of large-scale labeled image datasets; e.g., ImageNet [40], Places[54], Places2 [53] and COCO [30]. Traditional pattern recognition methods [1], [3], [5], [4], [51], [55], [48], [35], [26], [34], [23], [21], [28], [18], [6], [7], [8], [9], [10], [11], [37], [39], [38], [29] fail to provide feasible solutions to these large-scale datasets. In contrast, the CNN approach is more scalable and offers the state-of-the-art performance. Generally speaking, scene classification is more challenging than object classification due to higher ambiguity in scene labels. To boost the performance of scene classification, it is essential to have a deeper understanding on confusing classes and offer a solution to address this inherent ambiguity. We will use the Places scene image dataset as our main application focus throughout the paper.

This work has several major contributions. First, we develop a simple yet systematic way to identify confusing classes. Being inspired by the RECOS model for CNNs in [25], we call the filter weights at the last stage of the trained CNN the "scene anchor vector (SAV)". An SAV points to a cluster of images as shown in Fig. 1. If two SAVs have a smaller angle, we see overlapping image clusters which lead to a set of confusing classes and call it a confusion set. To the best of our knowledge, this is the first work that uses the SAV concept for confusing class identification in the Places dataset. Second, we propose a method to enhance the scene classification performance among confusing classes automatically. We merge images in a confusion set and, then, split them in an unsupervised manner to create multiple subsets. This is called the "automatic subset clustering (ASC)" process since it is done without supervision. Each of these subsets contains scene images of strong visual similarity. Then, we can zoom into each subset, select proper features and train another classifier. Here, we train a random forest (RF) classifier within each subset. The ASC/RF scheme can be added on top of any existing scene-classification CNN (called a baseline method) as a post-processing module to allow better scene classification with little additional training effort. Finally, we show by extensive experimental results that the ASC/RF module can offer a significant performance gain over the baseline method.

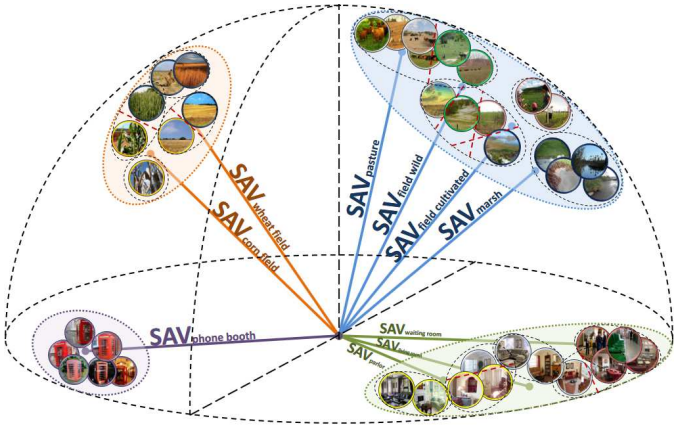

Fig. 1. Illustration of the scene anchor vectors (SAVs), which are the filter weights at the last stage of a CNN associated with certain scene labels trained by the Places dataset. One image cluster is associated with an SAV. The four blue SAVs (pasture, field wild, field cultivated and marsh) are close to each other in their angles and, as a result, they form one confusion set. Three other confusion sets are also shown in the figure.

The rest of this paper is organized as follows. Related previous work is reviewed in Sec. II. The SAV concept is introduced and used to determine the confusing scene classes

for the Places Dataset in Sec. III. Afterwards, the ASC process to create multiple scene subsets from a confusion set is presented and the integrated ASC/RF solution is described in Sec. IV. Extensive experimental results are shown to support our proposed solution in Sec. V. Finally, concluding remarks are given in Sec. VI.

II. Related Work

One trend in CNNs is to increase the number of convolutional layers as evidenced by the evolution from the LeNet [27], AlexNet [24], VGG [43] to the ResidueNet [20]. However, this is accompanied by the growing model complexity which in turn creates a challenging optimization problem in training a network. In the scene classification field, there is a great amount of work focusing on solving the optimization problem. The state-of-the-art approach [41] is one of the leaders in this group. It gets the first place in the Places2 [53] challenge using Relay Backpropagation (Relay BP). It encourages the propagation of effective information through the network in the training stage. However, this approach demands a large amount of time and engineering effort during the training process.

Another research effort, *e.g.* [15], [12], [32], exploit trained CNN models for better scene classification performance by enhance feature representations. Dixit *et al.* [15] obtained fisher embeddings from local predictions in an image to improve the learned features from the trained VGG network. Cheng *et al.* [12] learned multi-level sparselets to explore the discriminative information hidden in the learned network neurons. Liu *et al.* [32] defined cross-layer pooling schemes to consider image patterns in different scales. Although these methods claim lower complexity as compared with networks of growing layers, they usually provide little gain over the performance of the referenced network.

As a combination of these two approaches, researchers also considered hybrid models by extracting features from a trained CNN and conduct another minor deep neural network trainings to boost the performance. For example, Perronnin and Larlus [36] cascaded the learned fisher vector representation from a trained CNN and new fully connected layers to form a hybrid network.

Despite the aforementioned efforts, the scene classification field encounters a major problem of class ambiguity nowadays. That is, the large intra-class variation and inter-class confusion are the main causes that prevent further improvement of scene classification performance. This challenge is not well addressed by any of the previous work. This observation motivates us to explore confusing scene classes in the Places dataset to gain insights and develop a solution.

In the object classification field, several methods were proposed to use object class relationships to boost object classification performance through confusion analysis. Early work in [22], [33], [46] defined hierarchies to enhance the object classification performance using shallow models. Zweig and Weinshall [56] adopted an ensemble of object classifiers for different class hierarchies to improve the performance. With the WordNet-based distance, Fergus *et al.* [17] defined shared labels of object categories. In [52], an object hierarchy is used to define loss functions. Researchers also tried to generate the hierarchy adaptively to divide a general classification problem to more specific ones to improve the overall performance. Examples include [2], [14], [19], [44], [31], [16].

More recently, with the CNN-based classifier, attention has been shifted to hierarchical analysis for CNN performance enhancement. A tree-structured transfer learning approach was proposed in [45] to boost the CNN performance. In [13], images were relabeled and a modified CNN with hierarchical and exclusive graphical models were used to improve the performance. Xiao *et al.* [47] used a CNN-based incremental learning method to leverage the category hierarchy and keep improving the model when new image data are available. In [49], coarse and fine categories were first defined by analyzing the confusing matrix, which was obtained from the CNN classification results, and a modified hierarchical CNN model was trained to improve the overall object classification. These efforts show the importance of incorporating scene class confusions and hierarchies. However, none of these solutions have been extended to a large-scale scene classification problem such as Places.

Being inspired by the work in resolving confusion of classifying object classes and the RECOS model for a CNN in [25], we propose to use the inherent information in a trained CNN to determine sets of confusing scene classes. Furthermore, we propose a solution to enhance the scene classification performance within each confusion set. These two topics are detailed in the next two sections.

III. Confusion Set Analysis

A. Insights into SAVs

Anchor vectors are filter weights in the intermediate layers of a CNN. This term was coined by Kuo in [25] due to their unique role in a CNN. This role was explained using the "REctified COrrelations on a Sphere (RECOS)" model in [25]. In a trained CNN, an anchor vector behaves like a centroid of a cluster of input vectors in the corresponding layer.

Here, we pay special attention to the last stage filters of a trained CNN model, which take the the last fully connected (FC) layer as the filter input and the output layer as the filter output. In the current context, the decision is a scene label. We plot several representative final-stage anchor vectors in Fig. 1. Being different from the anchor vectors in previous

layers, anchor vectors of the last stage provide the ultimate feature representation of the input image, and the CNN has to make decision based on these feature vectors. Because of this special physical meaning, they are called the scene anchor vectors (SAVs).

Each SAV behaves like a centroid of scene images, and a great majority of them around an SAV share the same scene label. That is, through the convolution (or projection) operation, they can provide the largest projection value with respect to their target SAV and, as a result, they are classified to the desired scene accurately. In the Places dataset, we have 205 classes. Thus, a CNN model trained by the Places dataset should have 205 SAVs and each of them corresponds to a scene class.

As explained earlier, there is one image cluster associated with each SAV. By zooming into each cluster, we find that it contains multiple sub-clusters. It is interesting to show images from sub-clusters to gain further insights. Examples are given in Fig. 2, which contains 4 sub-clusters (or blocks) partitioned by thick black lines. All of them are pointed by the SAV corresponding to the "aqueduct" class. Each block contains 4 sample images. We observe strong visual similarities of images in the same block.

The top two blocks have images only from the "aqueduct" class. They look different from each other due to the opposite (left versus right) directions of their vanishing lines. This shows intra-class variations in the "aqueduct" class. On the other hand, images in the bottom two blocks are from different classes. They also share similar visual appearance within each block. For example, the two "boardwalk" images and one "bridge" image in the bottom-left corner all have a bridge grid pattern which also appears in the "aqueduct" image of the same block. Besides, the "viaduct" and the "bridge" images in the bottom-right corner have the arch-like supporting structure which are visually similar to the two "aqueduct" images of the same block. These two representative subsets provide excellent examples for inter-class ambiguity in the set of images pointed by the "aqueduct" SAV. Apparently, identifying the confusing image set and the corresponding subsets is critical to the understanding of the source of confusion.

B. Confusion Set Identification

We can build a graph to represent the relationship between scene classes in the Places dataset. Each node in the graph corresponds to a scene class. Every two nodes have an edge whose weight is determined by the cosine function of the angle between their corresponding SAVs:

$$\cos[\theta(\mathbf{a_i}, \mathbf{a_j})] = \frac{\mathbf{a_i}^T \mathbf{a_j}}{|\mathbf{a_i}||\mathbf{a_j}|} \tag{1}$$

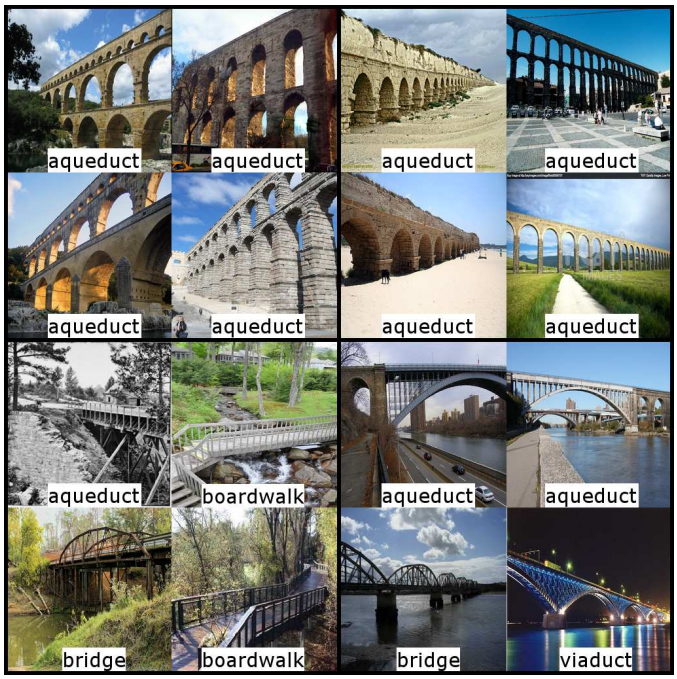

Fig. 2. Illustration of images in four subsets pointed by the "aqueduct" SAV.

where $\mathbf{a_i}$ and $\mathbf{a_j}$ are two SAVs. The above process generates a fully connected graph. We can simplify the graph by a thresholding step:

$$E_{\mathbf{a_i},\mathbf{a_j}} = \begin{cases} \cos[\theta(\mathbf{a_i}, \mathbf{a_j})] & \text{if } \theta(\mathbf{a_i}, \mathbf{a_j}) \geq T \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

where $0 < T < 1$ is a preselected threshold value. As shown in the above equations, the two nodes are more correlated and thus connected by an edge in the graph if their angle is smaller or their correlation is stronger. Otherwise, they are disconnected. Note that we can convert a densely-connected graph to a sparsely-connected graph by using a larger value of T. Fixing a value of T, we can use a Normalized-Cut algorithm [42], [51] to further simplify the graph and find cliques in the graph. Each resulting clique is called a confusion set that consists of several confusing classes that are challenging to separate for the trained network.

Although it is possible to derive confusion sets directly from the classification confusion matrix, the SAV-based confusion set identification method has two major advantages over the alternative method. First, we need a sufficiently large number of image samples to obtain the confusion matrix to avoid data bias. In contrast, this is not a problem for SAV since SAVs are directly derived from the trained network. They are actually statistical results of a large number of training images, and no further classification process is needed. We will show this advantage in Sec. V. Second, the confusion matrix method mainly offers the top-1 confusion errors between classes. It would be too complicated to track the top-n ($n > 1$) classfication errors. On the other hand, we

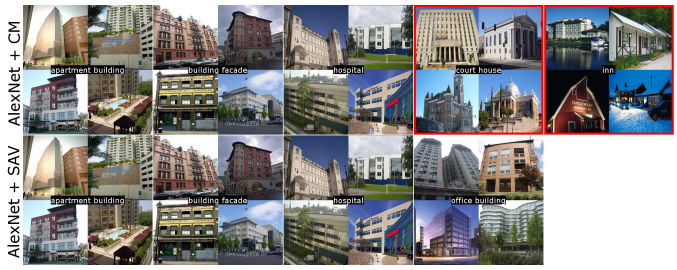

Fig. 3. Confusion set identification results comparison using the confusion matrix (CM) method and the SAV-based method.

can determine the top-n error rates easily by measuring the distance between an image sample to different SAVs.

Fig. 3 gives an excellent proof of the superiority of the SAV-based method. In the figure, we show two confusion sets for the Places dataset based on the AlexNet CNN. They are obtained by the confusion matrix (CM) method and the SAV-based method, respectively. For the CM method, the confusion set contains five confusing classes (from top left to top right: apartment building, building facade, hospital, court house and inn.) Most images in the last two classes, which are highlighted by a red box, are not visually similar. In contrast, the confusion set derived by the SAV-based method contains four confusing classes (from bottom left to bottom right: apartment building, building facade, hospital and office building.) We see that the first three confusing sets in these two methods are identical. However, the confusion set of the SAV-based method does not have the court house and inn classes but the office building class. Clearly, images in the office building class are more visually similar to images in the other three classes. This demonstrates the power of the proposed SAV-based method in confusion set identification.

IV. Boosting Performance in Confusion Sets

A. Automatic Subset Clustering (ASC)

We show how to identify sets of confusing classes using SAV correlations in Sec. III-B. To further resolve confusions within a confusion set, we propose to use an "automatic subset clustering (ASC)" algorithm to cluster images into subsets as shown in Fig. 2, where the number of subsets is automatically decided in the clustering process. The ASC algorithm is an *unsupervised* splitting process based on the k-means clustering idea with $k = 2$. The binary clustering algorithm in [51] is adopted in our implementation. The ASC algorithm is also a recursive clustering process, where each node is split into two child nodes until one of the stopping criteria is met.

The stopping criteria are used to decide whether to split a set of images in one node into two child nodes. We propose two criteria and consider them in a sequential order. First, we check whether the images in a node are from the same scene

Fig. 4. Illustration of the ASC process, where the border color of each image represents its class type (Red: "field wild"; Purple: "field cultivated"; Green: "pasture"; Blue: "marsh") and the border color of each image block indicates its subset property (Pink: "the subset needs a further split."; Light green: "the subset is pure and does not need a further split."; Yellow: "the subset has a sufficiently small variance value and does not need a further split.")

class or not. If so, this node is a "pure" node and no further split is needed. Otherwise, the node is an "impure" one. For an impure node, we check it using the second criterion. It is based on the total variance of images in the node, which can be computed by

$$\overline{V}_{anchor} = \frac{1}{N_c}\sum_{i=0}^{N_c}\frac{1}{M_i}\sum_{j=0}^{M_c} E_{\mathbf{x_j},\mathbf{a_i}} \tag{3}$$

where N_c is the number of class types, M_i is the number of images in class i in the node, and $E_{\mathbf{x_j},\mathbf{a_i}}$ is the correlation between an image and an anchor vector of a class, which follows the definition given in Eq. 2. If $\overline{V}_{anchor}$ is smaller than a pre-selected threshold value denoted by T_v, the split process is stopped. This means that images in this node are visually similar, yet they belong to different classes. In Fig. 4, the two image blocks with the yellow border are good examples for this case. No further splitting is needed since it does not help much. This is truly a challenging subset and we need to find another solution to resolve the confusion. This will be detailed in Sec. IV-B. On the other hand, if $\overline{V}_{anchor}$ value is larger than T_v, it is a node with the pink border in the figure. Further splitting is needed.

A representative ASC process is illustrated in Fig. 4. The root node at the top level represents a confusion set with the pink border. It contains images from 4 classes (i.e., "pasture" and "field wild", "field cultivated" and "marsh"). We split it into two child nodes. Images in the left child node with the yellow border do not satisfy the first stopping criterion but the second one. We see that these images are visually similar yet with different class labels. The right child node is further split into two child nodes at the 3rd level. The right node in the 3rd level is a pure one (with the green

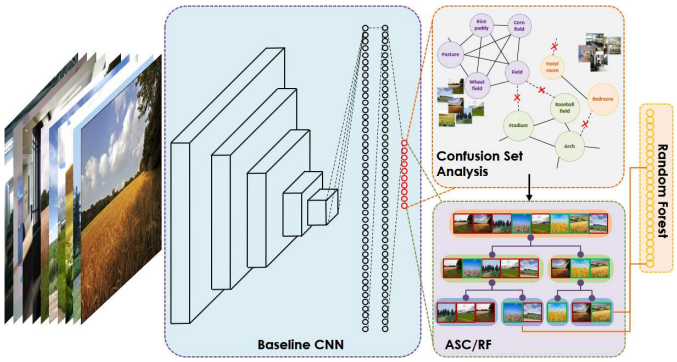

Fig. 5. A post-processing module that integrates the ASC process and the random forest (RF) classifier to boost the performance of a baseline CNN in confusing sets.

border) which satisfied the first stopping criterion. The left node (with the pink border) is further split into two child nodes. The right node in the 4th level is a pure one (with the green border) while the left node in the 4th level satisfies the 2nd stopping criterion (an impure node with a small variance).

For this particular example, we obtain 4 leaf nodes in the binary hierarchical tree - two pure subsets and two impure subsets at the end of the ASC process. For the two pure subsets, we have images from the "field cultivated" and the "marsh" classes, respectively. We see visually similar images that contain humans and crops in images of the "field cultivated" class and water surface and green plants in images of the "marsh" class. This demonstrates the power of ASC to identify distinguishable pure image subsets in an unsupervised manner. It is worthwhile to point out that the two impure subsets also contain visually similar images. For example, animals and fields are observed from images of the yellow-border node in the second level, and wide landscape views are observed from images of the yellow-border node in the fourth level. This demonstrates the power of combining three components: 1) the use of feature vectors at the last FC layer, 2) the adoption of the SAV concept in clustering image samples, and 3) the use of the ASC process to split a confusion set into multiple subsets.

B. Post-Processing by Integrated ASC/RF

The understanding of the source of confusion as discussed in Sec. III-B and Sec. IV-A allows us to measure the capability of a baseline network and identify the challenging samples in the Places dataset. Furthermore, based on the understanding, we propose a post-processing technique to boost the performance in confusion sets for the baseline method.

Fig. 5 presents an overview of the proposed system, where the left part includes a set of input images, the middle part shows a baseline CNN and the right part is the proposed post-processing module that consists of two components: confusion set analysis and the ASF/RF scheme. The latter will be presented below. In the training stage, given a baseline CNN for the Places set (e.g. the trained AlexNet or VGG16), we first determine confusion sets by analyzing SAV correlations as discussed in Sec. III. Then, for each confusion set, we conduct the ASC process to cluster images into subsets as described in Sec. IV-A. Finally, we train a multi-class RF classifier in each impure subset to enhance the classification performance.

The choice of the RF classifier is justified below. The feature vectors obtained by the baseline CNN have an outstanding discriminant power as shown in the example in Fig. 4. We would like to re-use the feature vector as much as possible in each impure node. The RF classifier contains a large number of decision trees and it can select the most discriminant feature in making decision in a sequential manner. Thus, we use the RF classifier to mine minor differences between confusing classes and images furthermore.

In the testing stage, we first obtain the feature vector of a test image using the baseline CNN. If it is located in a confusion set, we assign it to a leaf node in the hierarchical tree by traversing the tree created by the ASC process for the particular confusion set. If the subset is pure, we simply output class label associated with that node. Otherwise, it is an impure node and we run the RF classifier trained by images in that node to provide a predicted class label.

V. Experimental Results

In this section, we conduct experiments to show the advantages of the two proposed methods; namely, the confusion set analysis in Sec. III and the post-processing ASC/RF module in Sec. IV. We first explain the experimental settings and, then, present the performance gain achieved by each single method. Finally, we will show the overall performance gain of the whole system consisting of both methods.

A. Experimental Settings

Dataset. We evaluate the proposed methods on the Places [54] dataset. It contains 205 categories of scene images. There are 2,469,373 images in total, where 2,448,873 images are used as training images and 20,500 images are used as testing images. The averaged number of images for the training set is more than 10,000 per class while that for the testing set is 100 per class.

Baselines and features. We present experimental results of the proposed methods over two baseline CNNs, AlexNet and VGG16 [54], which are applied to the Places dataset. In the proposed system, confusion set identification and ASC/RF are two individual modules that both contribute to

TABLE I
COMPARISON OF AVERAGED PRECISION (AP) AND STANDARD DEVISIONS (STD) OF 5-FOLD CLASSIFICATIONS VALIDATION USING CM AND SAV.

Methods	AP	STD
AlexNet/CM	50.01%	1.34%
AlexNet/SAV(Ours)	**51.53%**	**0.41%**
VGG16/CM	58.76%	2.07%
VGG16/SAV(Ours)	**60.94%**	**0.68%**

performance enhancement. For the confusion set identification module, we use the filter weights from the FC8 layer (the last-stage constitutional layer) of a CNN as the SAV in identifying confusion sets. For the ASC/RF post-processing module, we adopt the layer output from FC7, which is the second-to-the-last layer in a trained CNN.

B. CM versus SAV for Confusion Set Identification

We focus on two aspects in the comparison of the CM method and the SAV-based method for confusion set identification.

Data dependency. The SAV-base method is determined by the CNN training data only. Once a CNN is completely trained, its SAVs are fixed. Then, one can conduct the confusion analysis without any additional data. In contrast, the CM method does need additional data to test the classification performance of the trained CNN. If we use the CNN training data for this purpose, the results tend to be bias. Since the trained CNN will provide biased results. As a consequence, we need a set of new training data. For this reason, we use 5-fold validations to evaluate the performance of the two methods, where 80% of the data were used to train the CNN and 20% of the data were used to obtain the CM. Then, we use the test dataset to evaluate the performance.

To compare the robustness of the two methods, we present the averaged precision (AP:averaged top-1 classification accuracy) and the corresponding standard devisions (STD) in TABLE I. Clearly, the SAV-based method has a larger AP and a smaller STD than the CM method for both AlexNet and VGG16.

Confusion set transferability. It is interesting to study the performance by transferring the confusion set knowledge learned by one CNN to the other CNN. For example, we compute the confusion set by analyzing the AlexNet using either the CM or the SAV-based method and apply this knowledge to a VGG16 CNN. The results are given in Table II. By comparing the results in the first column of TABLE I and TABLE II, we can see the performance drop due to the knowledge transfer. The performance of the AlexNet with the CM method, drops from 50.01% to 49.01% while the performance of the AlexNet with the SAV-based method drops from 51.53% to 51.42%. The performance of the VGG16 with the CM method, drops from 58.76% to 58.01% while the performance of the VGG16 with the SAV-based method drops from 60.94% to 60.77%. It is clear that the SAV-based method has a more robust performance against the confusion set knowledge transfer.

TABLE II
COMPARISON OF TOP-1 AND TOP-5 CLASSIFICATION PERFORMANCE BY TRANSFERRING THE CONFUSION SET KNOWLEGE LEARNED FROM ONE CNN TO ANOTHER CNN.

Methods	Top-1	Top-5
AlexNet/CM(VGG16)	49.01%	80.15%
AlexNet/SAV(VGG16)	51.42%	81.27%
VGG16/CM(AlexNet)	58.01%	86.23%
VGG16/SAV(AlexNet)	60.77%	88.12%

TABLE III
CLASSIFICATION PERFORMANCE COMPARISON USING ASC ALONE, RF ALONE AND JOINT ASC/RF.

Methods	Top-1	Top-5
AlexNet + ASC	50.62%	80.80%
AlexNet + RF	50.10%	80.03%
AlexNet + ASC/RF	51.31%	81.15%
VGG + ASC	59.37%	87.67%
VGG + RF	59.32%	87.12%
VGG + ASC/RF	60.84%	88.06%

C. Evaluation of ASC/RF Post-processing

In this subsection, we would like to evaluate three combinations (i.e., ASC alone, RF alone and joint ASC/RF) to understand their contributions more clearly.

Automatic Subset Clustering (ASC). In the joint ASC/RF module, an RF classifier is used to make further judgment in mixed subsets after ASC. As an alternative, we can simply output the label of the class that has the largest number of samples. This is a good test for the ASC scheme since samples of different classes can be well separated by a good ASC scheme. We compare the top-1 and top-5 classification performance of three post-processing techniques in TABLE III. As compared with the joint ASC/RF post-processing, the performance drops in the top-1 accuracy of the ASC alone are 0.69% and 1.47% for AlexNet and VGG16, respectively.

RF confusion set classifiers. When the ASC procedure is removed, we obtain a larger set of confusing classes. We can still train a RF classifier for that particular confusion set to boost the classification performance. The results are shown in TABLE III. As compared with the joint ASC/RF, we see significant performance drops, namely; 1.21% (top-1) and 1.12% (top-5) for AlexNet and 1.52% (top-1) and 0.94% (top-5) for VGG16. These results show the necessity of conducting the ASC procedure before the RF classification.

TABLE IV
COMPARISON OF TOP-1 AND TOP-5 CLASSIFICATION ACCURACIES OF ALEXNET, VGG16 AND THEIR ENHANCED SOLUTIONS AGAINST THE PLACES DATASET.

Methods	Top-1	Top-5
AlexNet	49.805%	80.244%
AlexNet (Enhanced)	**51.307%**	**81.148%**
VGG16	58.526%	86.731%
VGG16 (Enhanced)	**60.843%**	**88.061%**

D. Overall Performance Improvement

We compare the classification performance of the two baseline CNNs and their enhanced versions, which include both SAV-based confusion set identification and joint ASC/RF post-processing, in TABLE IV. We see from the table that the fully integrated system outperforms its baseline CNN by a significant margin. For the AlexNet, we see 1.5% (top-1) and 0.9% (top-5) improvement. For the VGG-16, we observe 2.3% (top-1) and 1.3% (top-5) improvement. These gains are impressive by considering the fact that we do not perform any additional CNN training as done in previous work, *e.g.*, [49], [36], [50], [41]. Furthermore, the proposed methodology can be applied to any baseline CNN for further performance gain.

It is worthwhile to emphasize that the Places dataset is well known for its challenging scene confusion. The large performance gain in the top-1 classification accuracy demonstrates the power of the proposed solution in resolving inter-class confusion. The lower performance gain in the top-5 classification accuracy can be explained by the fact that the ASC/RF post-processing module focuses on resolving confusion within a confusion set, yet many confusion sets contain less than 5 confusing scene classes.

To further demonstrate the power of the proposed confusion set resolution scheme, we list the improvement of top-1 and top-5 classification accuracies in TABLE V and TABLE VI, respectively, in several representative confusion sets. Generally speaking, a confusion set with a smaller number of confusing classes has a higher top-1 accuracy improvement. On the other hand, a confusion set with a larger number of confusing classes has a higher top-5 accuracy improvement. This observation can be explained as follows.

If the confusion set size is large in terms of the number of confusing classes, there exist severe ambiguities between images from different classes in the set. This makes the top-1 classification enhancement more challenging. On the other hand, there are more learning samples in the confusion set, leading to more reasonable guesses to the ground-truth label. Thus, we see more improvement in the top-5 accuracy. If the confusion set size is small, there is less confusion but a limited number of training samples for the ASC/RF module.

TABLE V
COMPARISON OF THE TOP-1 CLASSIFICATION ACCURACY FOR VGG16 AND ENHANCED VGG16 IN FOUR REPRESENTATIVE CONFUSION SETS.

Confusion Sets	Top-1 Accuracy	
	VGG16	**Ours**
fairway, golf course	60.34%	**75.21%**
galley, kitchen, kitchenette, pantry, restaurant kitchen	69.74%	**80.14%**
botanical garden, orchard, formal garden, herb garden, cottage garden, topiary garden, vegetable garden	63.01%	**72.01%**
apartment building, skyscraper, building facade, fire escape, hospital, office building	63.18%	**71.70%**

TABLE VI
COMPARISON OF THE TOP-5 CLASSIFICATION ACCURACY FOR VGG16 AND ENHANCED VGG16 IN FOUR REPRESENTIVE CONFUSION SETS.

Confusion Sets	Top-5 Accuracy	
	VGG16	**Ours**
banque hall, bar, beauty salon, bakery shop, cafeteria, coffee shop, dining room, dinette home, food court, ice cream parlor, restaurant, restaurant patio	81.22%	**92.34%**
crevasse, iceberg, igloo, ice skating rink outdoor, mountain snowy, ski resort, ski slope, snowfield	80.11%	**89.67%**
ballroom, game room, office, home office, music studio, reception, stage indoor, television studio	85.68%	**90.75%**
aqueduct, boardwalk, bridge, pavilion, rope bridge,viaduct	92.31%	**95.28%**

Finally, we present four test images and show the process of confusion resolution in Fig. 6. The test image in the first column is from the “game room” class and it is assigned to a pure subset in the ASC process, which leads to an efficient correct labeling. The test image in the last column is from the “court yard” class. It is assigned to a confusion set at the first level. Then, it goes to a mixed subset consisting of two confusing classes - court yard and picnic area. It is eventually classified to the correct class using the RF classifier. The probability distribution of confusing classes along the classification path in the ASC/RF system for each test image is clearly demonstrated. For all four cases, we are able to get the correct result. It is interesting to point out that the coast image in the second column could be assigned to the “islet” class at the first level since the latter has a higher probability. However, as the image moves along the path, the probability of the coast class becomes the highest.

VI. CONCLUSIONS AND FUTURE WORK

The Places dataset is the most challenging scene classification dataset today for its challenging intra-class

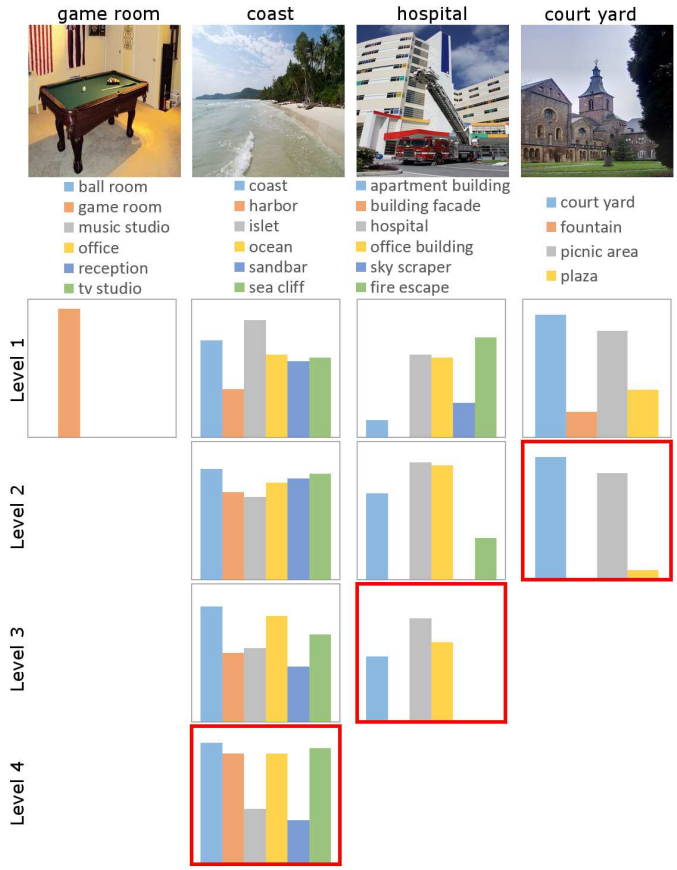

Fig. 6. Studies on the classification path of four correctly labeled images in the Places dataset, where each column shows a test case. The first row gives the distribution of confusing classes in the assigned assigned confusing set. The red box indicates a decision is made using a RF classifier based on samples at that node.

variation and inter-class ambiguities. In this work, we first proposed an SAV-based method to identify confusing sets. Then, to enhance the classification performance within a confusion set, we proposed an ASC/RF post-processing module. Extensive experiments were shown to demonstrate the significant gains of the proposed enhancement schemes.

There are several possible extensions of the current work in the future. First, we are interested in applying the proposed methodology on top of more advanced networks such as ResNet. Second, it is promising to design a soft assignment scheme that allows a test image to have multiple subset labels in the ASC procedure. Then, we can conduct a weighted average to boost the classification accuracy. Finally, it appears to be meaningful to extend the confusing class identification technique to multi-level confusion hierarchies to obtain even higher classification performance improvement in confusion sets.

REFERENCES

[1] M. M. Ali, M. B. Fayek, and E. E. Hemayed. Human-inspired features for natural scene classification. *Pattern Recognition Letters*, 34(13):1525–1530, 2013.
[2] H. Bannour and C. Hudelot. Hierarchical image annotation using semantic hierarchies. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 2431–2434. ACM, 2012.
[3] A. Bolovinou, I. Pratikakis, and S. Perantonis. Bag of spatio-visual words for context inference in scene classification. *Pattern Recognition*, 46(3):1039–1053, 2013.
[4] A. Bosch, A. Zisserman, and X. Muoz. Scene classification using a hybrid generative/discriminative approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(4):712–727, 2008.
[5] A. Bosch, A. Zisserman, and X. Muoz. *Scene classification via pLSA*, pages 517–530. Springer, 2006.
[6] C. Chen, Y. Ren, and C.-C. J. Kuo. Large-scale indoor/outdoor image classification via expert decision fusion (edf). In *Asian Conference on Computer Vision*, pages 426–442. Springer International Publishing, 2014.
[7] C. Chen, Y. Ren, and C.-C. J. Kuo. Big visual data analysis: scene classification and geometric labeling, 2016.
[8] C. Chen, Y. Ren, and C.-C. J. Kuo. Global-attributes assisted outdoor scene geometric labeling. In *Big Visual Data Analysis*, pages 93–120. Springer, 2016.
[9] C. Chen, Y. Ren, and C.-C. J. Kuo. Indoor/outdoor classification with multiple experts. In *Big Visual Data Analysis*, pages 23–63. Springer, 2016.
[10] C. Chen, Y. Ren, and C.-C. J. Kuo. Outdoor scene classification using labeled segments. In *Big Visual Data Analysis*, pages 65–92. Springer, 2016.
[11] C. Chen, Y. Ren, and C.-C. J. Kuo. Scene understanding datasets. In *Big Visual Data Analysis*, pages 7–21. Springer, 2016.
[12] G. Cheng, J. Han, L. Guo, and T. Liu. Learning coarse-to-fine sparselets for efficient object detection and scene classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1181, 2015.
[13] J. Deng, N. Ding, Y. Jia, A. Frome, K. Murphy, S. Bengio, Y. Li, H. Neven, and H. Adam. Large-scale object classification using label relation graphs. In *European Conference on Computer Vision*, pages 48–64. Springer, 2014.
[14] J. Deng, S. Satheesh, A. C. Berg, and F. Li. Fast and balanced: Efficient label tree learning for large scale object recognition. In *Advances in Neural Information Processing Systems*, pages 567–575, 2011.
[15] M. Dixit, S. Chen, D. Gao, N. Rasiwasia, and N. Vasconcelos. Scene classification with semantic fisher vectors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2974–2983, 2015.
[16] J. Dong, Q. Chen, J. Feng, K. Jia, Z. Huang, and S. Yan. Looking inside category: subcategory-aware object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(8):1322–1334, 2015.
[17] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *European Conference on Computer Vision*, pages 762–775. Springer, 2010.
[18] D. Gokalp and S. Aksoy. Scene classification using bag-of-regions representations. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.
[19] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015.
[21] H. Jgou, M. Douze, C. Schmid, and P. Prez. Aggregating local descriptors into a compact image representation. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3304–3311. IEEE, 2010.
[22] Y. Jia, J. T. Abbott, J. Austerweil, T. Griffiths, and T. Darrell. Visual concept learning: Combining machine vision and bayesian generalization on concept hierarchies. In *Advances in Neural Information Processing Systems*, pages 1842–1850, 2013.
[23] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 923–930. IEEE, 2013.
[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification

with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.
[25] C.-C. J. Kuo. Understanding convolutional neural networks with a mathematical model. *Journal of Visual Communication and Image Representation*, 41:406–413, 2016.
[26] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 2, pages 2169–2178. IEEE, 2006.
[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
[28] L.-J. Li, R. Socher, and L. Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2036–2043. IEEE, 2009.
[29] S. Li, S. Purushotham, C. Chen, Y. Ren, and C.-C. J. Kuo. Measuring and predicting tag importance for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.
[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollr, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755. Springer, 2014.
[31] B. Liu, F. Sadeghi, M. Tappen, O. Shamir, and C. Liu. Probabilistic label trees for efficient large scale image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 843–850, 2013.
[32] L. Liu, C. Shen, and A. van den Hengel. The treasure beneath convolutional layers: Cross-convolutional-layer pooling for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4749–4757, 2015.
[33] M. Marszalek and C. Schmid. Semantic hierarchies for visual object recognition. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–7. IEEE, 2007.
[34] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175, 2001.
[35] S. N. Parizi, J. G. Oberlin, and P. F. Felzenszwalb. Reconfigurable models for scene recognition. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2775–2782. IEEE, 2012.
[36] F. Perronnin and D. Larlus. Fisher vectors meet neural networks: A hybrid classification architecture. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3743–3752, 2015.
[37] Y. Ren. *Techniques for vanishing point detection*. University of Southern California, 2013.
[38] Y. Ren, C. Chen, S. Li, and C.-C. J. Kuo. Gal: A global-attributes assisted labeling system for outdoor scenes. *Journal of Visual Communication and Image Representation*, 42:192–206, 2017.
[39] Y. Ren, S. Li, C. Chen, and C.-C. J. Kuo. A coarse-to-fine indoor layout estimation (cfile) method. In *Asian Conference on Computer Vision*, pages 36–51. Springer, Cham, 2016.
[40] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.
[41] L. Shen, Z. Lin, and Q. Huang. Relay backpropagation for effective learning of deep convolutional neural networks. *arXiv preprint arXiv:1512.05830*, 2015.
[42] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
[44] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros. Unsupervised discovery of visual object class hierarchies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.
[45] N. Srivastava and R. R. Salakhutdinov. Discriminative transfer learning with tree-based priors. In *Advances in Neural Information Processing Systems*, pages 2094–2102, 2013.
[46] N. Verma, D. Mahajan, S. Sellamanickam, and V. Nair. Learning hierarchical similarity metrics. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2280–2287. IEEE, 2012.
[47] T. Xiao, J. Zhang, K. Yang, Y. Peng, and Z. Zhang. Error-driven incremental learning in deep convolutional neural network for large-scale image classification. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 177–186. ACM, 2014.
[48] L. Xie, J. Wang, B. Guo, B. Zhang, and Q. Tian. Orientational pyramid matching for recognizing indoor scenes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3734–3741. IEEE, 2014.
[49] Z. Yan, V. Jagadeesh, D. Decoste, W. Di, and R. Piramuthu. Hd-cnn: hierarchical deep convolutional neural network for image classification. In *International Conference on Computer Vision (ICCV)*, volume 2, 2015.
[50] S. Yang and D. Ramanan. Multi-scale recognition with dag-cnns. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1215–1223, 2015.
[51] S. X. Yu and J. Shi. Multiclass spectral clustering. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 313–319. IEEE, 2003.
[52] B. Zhao, F. Li, and E. P. Xing. Large-scale category structure aware image categorization. In *Advances in Neural Information Processing Systems*, pages 1251–1259, 2011.
[53] 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.
[54] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.
[55] L. Zhou, Z. Zhou, and D. Hu. Scene classification using a multi-resolution bag-of-features model. *Pattern Recognition*, 46(1):424–433, 2013.
[56] A. Zweig and D. Weinshall. Exploiting object hierarchy: Combining models from different category levels. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.