The Thirty-Second AAAI Conference
on Artificial Intelligence (AAAI-18)

# DF$^2$Net: A Discriminative Feature Learning and Fusion Network for RGB-D Indoor Scene Classification

**Yabei Li,**[1,2] **Junge Zhang,**[1,2] **Yanhua Cheng,**[3] **Kaiqi Huang,**[1,2,4] **Tieniu Tan**[1,2,4]
[1]CRIPAC & NLPR, CASIA [2]University of Chinese Academy of Sciences [3]Tencent Wechat AI
[4]CAS Center for Excellence in Brain Science and Intelligence Technology
yabei.li@cripac.ia.ac.cn, {jgzhang, kqhuang, tnt}@nlpr.ia.ac.cn, breezecheng@tencent.com

## Abstract

This paper focuses on the task of RGB-D indoor scene classification. It is a very challenging task due to two folds. 1) Learning robust representation for indoor scene is difficult because of various objects and layouts. 2) Fusing the complementary cues in RGB and Depth is nontrivial since there are large semantic gaps between the two modalities. Most existing works learn representation for classification by training a deep network with softmax loss and fuse the two modalities by simply concatenating the features of them. However, these pipelines do not explicitly consider intra-class and inter-class similarity as well as inter-modal intrinsic relationships. To address these problems, this paper proposes a Discriminative Feature Learning and Fusion Network (DF$^2$Net) with two-stage training. In the first stage, to better represent scene in each modality, a deep multi-task network is constructed to simultaneously minimize the structured loss and the softmax loss. In the second stage, we design a novel discriminative fusion network which is able to learn correlative features of multiple modalities and distinctive features of each modality. Extensive analysis and experiments on SUN RGB-D Dataset and NYU Depth Dataset V2 show the superiority of DF$^2$Net over other state-of-the-art methods in RGB-D indoor scene classification task.

## Introduction

Scene classification is one of the basic problems in computer vision research. Recently, with the release of cost-affordable depth sensors, e.g. Kinect, which provide strong illumination and color invariant geometric cues, some intrinsic challenges in indoor scene classification such as various illumination, diverse objects and layouts are promising to be partially solved.

Compared with the standard object-centric image classification problem, the task of RGB-D indoor scene classification has several challenges. Firstly, obtaining robust representation for scene classification in single modality is difficult. To understand a scene, people not only recognize the objects in the scene, but also consider the correlations of the objects. As for indoor scenes, they are usually cluttered with diverse objects and various layouts, resulting in large intra-class variation and sever inter-class overlap. As we illustrated in Figure 1, the *classroom* has various views. Some



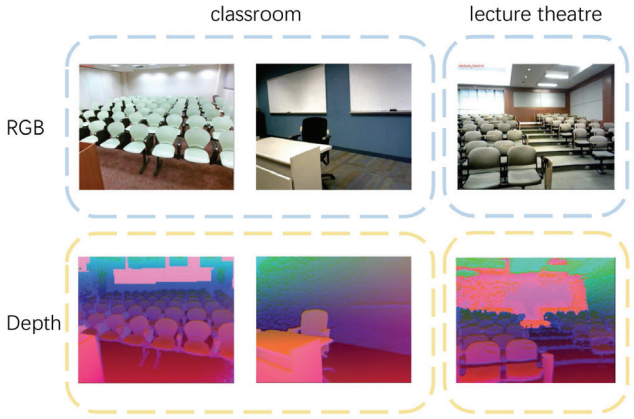


Figure 1: The difficulties of RGB-D indoor scene classification. 1) indoor scene images have large intra-class variation and small inter-class variation. 2) RGB and Depth image have semantic gaps. Sample images are from SUN RGB-D Dataset.

view of *classroom* is similar to other scene categories such as *lecture theatre*. Secondly, although it's an opportunity to utilize the additional depth cues to benefit indoor scene classification, there are large semantic gaps between the RGB and Depth modality. As shown in Figure 1, RGB image gives appearance cues while Depth image provides geometric priors. How to exhaustively use the complementary cues in the RGB and Depth modalities remains an open problem.

The ideal multimodal representation for RGB-D indoor scene ought to have small distances between the same class and large distances between the different classes, as shown in Figure 2(a). One of the most popular pipelines (Eitel et al. 2015) is to learn representation for RGB and Depth image with softmax loss separately and directly concatenate them, illustrated as in Figure 2(b). However, for scene images that have large intra-class variation and small inter-class variation, it's hard to obtain discriminative enough RGB and Depth representation. It leads to the concatenated multimodal representation far from ideal. To ease this situation, we design a discriminative feature learning network which can explicitly model the intra-class and inter-class similarity

constraint. The proposed pipeline is able to learn more discriminative representation for each modality and thus more discriminative concatenated multimodal representation. The effect can be shown in Figure 2(c).

Furthermore, simply concatenating RGB and Depth features does not effectively exploit the correlation between the two modalities. (Wang et al. 2015b) points out to learn consistency in RGB and Depth representation by optimizing a correlation term. It aims to close the distance of *holistic* embedded RGB and Depth representation of each image pairs. The pipeline can be illustrated in Figure 2(d). (Zhu, Weibel, and Lu 2016) improves the correlation term by considering label information. In spite of the effectiveness of them, we argue that enforcing the *holistic* representation for RGB and Depth to be correlated is not optimal for indoor scene classification. Although guaranteeing consistency between the two modalities is helpful to remove noises, it also removes the complementary information in RGB and Depth. For example, as illustrated in Figure 1, the RGB image uses *whiteboard* to help to represent the *classroom*, while the Depth image which emphasizes geometric cues has little information about the *whiteboard*. Enforcing the RGB and Depth representation to be wholly correlated will cause the representation focus on common information in RGB and Depth modalities (like the *chairs* and *tables* in *classroom*) and ignore distinctive information (such as *whiteboard*) for indoor scene in two modalities. For indoor scenes that need multiple contexts to be correctly recognized, representation that contain only consistency information in the two modalities is obviously not discriminative enough. Instead, in this paper, we construct a discriminative fusion network to learn both distinctive information and correlative information between the RGB and Depth modalities. Our framework can be illustrated as in Figure 2(e).

A novel Discriminative Feature Learning and Fusion Network (DF$^2$Net) is proposed with two-stage learning for RGB-D indoor scene classification. The contributions of this paper are threefold. 1) To learn more discriminative representation of each modality for indoor scene classification, a multi-task network is designed to explicitly encourage the intra-class compactness and inter-class separability in the first stage. 2) To fuse RGB and Depth representation effectively, a novel network is constructed in the second stage to learn both correlation and independent information in the two modalities. 3) We achieve the state-of-the-art results on both SUN RGB-D Dataset and NYU-Depth Dataset V2.

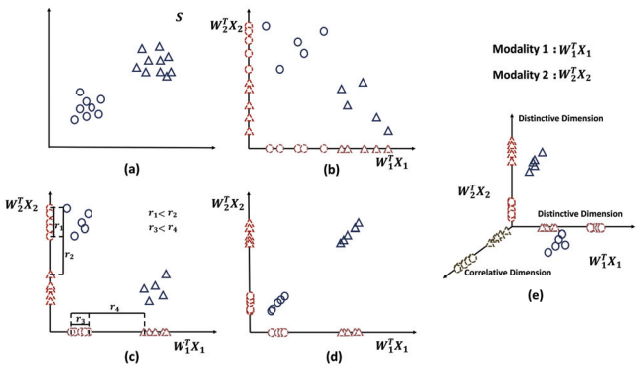


Figure 2: Different multimodal representation learning strategies. The "circle" and "triangle" represent different classes. The dark red, light red and blue samples illustrate the representation of modality1, modality2, and fused multimodal respectively. In e), each modality is represented as a plane with one correlation dimension and one distinctive dimension. a) Ideal discriminative features for RGB-D scene classification. b) Concatenating features from separately learned RGB and Depth representation directly. c) Effect of using more discriminative RGB and Depth representation for each modality. d) Enforcing the holistic RGB and Depth features to have correlation. e) Only enforcing part of RGB and Depth features to be correlated. The other parts of RGB and Depth features are encouraged to learn independent information separately.

## Related Work

**Scene Classification** Remarkable efforts have been invested to explore the discriminative representation for general scene. These works can be divided into three categories: 1) (Zhou et al. 2014) trains the Convolutional Neural Network (CNN) from scratch using collected large scale scene-centric images (Places Dataset). The trained CNN (Places CNN) can be employed to extract features for scene images. Though features extracted from Places CNN are more spatially diverse than that extracted from ImageNet pre-trained CNN, they are still too coarse to directly represent the indoor scene. 2) A lot of works use encoding methods such as Fisher Vector (FV) or Vector of Locally Aggregated Descriptors (VLAD) to combine features extracted from different scales and locations in image (Dixit et al. 2015; Arandjelovic et al. 2016; Wang et al. 2017). However, for indoor scene, some patches in scene images can be noisy for classification. Moreover, when the pipeline is plugged into CNN framework, the codebook selection and encoding procedure disables the whole framework to be trained jointly, resulting in the feature learning suboptimal. 3) Another popular pipeline employs objects and semantic parts in images to represent scene (Li et al. 2010; Wu et al. 2015; Wang et al. 2016; Bappy, Paul, and Roy-Chowdhury 2016). This pipeline can avoid some noisy information in scene images. Nevertheless, it needs to accurately detect objects, which by itself is quite difficult for cluttered indoor scene. Besides, not all objects in scene images are beneficial to classification. In this paper, we propose a new pipeline to better represent indoor scene for each modality, which can overcome the weaknesses in the three mainstream methods.

**RGB and Depth Image Fusion** The typical RGB and Depth fusion methods can be divided into three categories: 1) fusion at image level (Couprie et al. 2013); 2) fusion at feature level (Eitel et al. 2015; Cheng et al. 2016); 3) fusion at score level (Gupta et al. 2014; Cheng et al. 2015b; 2015a; 2017). The most related fusion methods to us are (Wang et al. 2015b) and (Zhu, Weibel, and Lu 2016), which also consider relationships between two modalities. (Wang et al. 2015a) directly enforces each RGB-D image pair to


Figure 3: The discriminative feature learning network for each modality. The input can be RGB or Depth image.

share part of features to learn common information and (Zhu, Weibel, and Lu 2016) requires the holistic representation of RGB-D image pair be discriminative and correlative simultaneously. Our method, however, is different from them. Designed for indoor scene classification, the proposed model aims to learn the distinctive and correlative parts of RGB and Depth representation separately. Besides, the correlations learned in our model are between not only RGB-D pairs but also random samples in RGB and Depth modalities. Moreover, both (Wang et al. 2015a) and (Zhu, Weibel, and Lu 2016) cannot be trained jointly while our framework can be optimized in an unified way.

**Deep Metric Learning** The deep metric learning can help to learn feature embedding that captures similarity. The most popular work of deep metric learning are contrastive embedding (Hadsell, Chopra, and LeCun 2006) and triplet embedding (Schroff, Kalenichenko, and Philbin 2015). They are widely used in many computer vision applications such as face recognition (Parkhi et al. 2015), image retrieval (Gordo et al. 2016), fine-grained classification (Wang et al. 2014). Some researchers also propose to use deep metric learning in multi-modal applications, but most of them focus on matching problem such as cross-modal retrieval (Yu et al. 2016; Huang et al. 2017). In this paper, we focus on multi-modal fusion problem, aiming to maximally utilize the complementary information in the two modalities.

## Methodology

In this section, the detailed framework will be described. The proposed DF$^2$Net is learned with two stages. In the first stage, the discriminative features for indoor scene are learned for RGB and Depth[1] modalities respectively. Then in the second stage, the learned discriminative features in the first stage are further exploited to learn independent and correlative representation for RGB-D indoor scene.

### Single-Modal Discriminative Feature Learning

The typical framework for object classification using CNN optimizes cross-entropy loss together with softmax (we call them together as softmax loss and denote it as $L_{softmax}$ in this paper) to learn features. It's simple and proved to be effective in many computer vision problems. However, in the softmax loss optimizing process, all images within the same category are mapped to a single point in feature space, which loses the intra-class variation. In this paper, we propose to explicitly model the similarity constrains. By optimizing a structured loss function, we are able to learn the manifold which can preserve the intrinsic intra-class variation and learn more discriminative representation for indoor scene classification (Cui et al. 2016; Zhang et al. 2016).

A multi-task discriminative feature learning network is proposed. The framework is shown in Figure 3. A fully connected layer is connected to the fc7 layer to reduce the feature dimension to 128. The structured loss connects to the 128-dimension feature to learn feature embedding. Here we learn the triplet embedding, which is supposed to have intra-class compactness and inter-class separability.

Denote the RGB or Depth inputs in a batch as $\{x_1, x_2, ..., x_N\}$ and their labels as $\{y_1, y_2, ..., y_N\}$. The feature embedding for them then can be represented as $\{f(x_1), f(x_2), ..., f(x_N)\}$. The structured loss we use is written as:

$$L_{structed} = \frac{1}{2\left|\mathcal{P}\right|} \sum_{(i,p)\in\mathcal{P}} \max(D_{ip}^2 - D_{in}^2 + \alpha, 0) \quad (1)$$

where, $(i,j) \in \mathcal{P}$ means $y_i = y_j$, $\alpha$ is the margin, $y_i = y_p$, $y_i \neq y_n$ and

$$D_{ij} = \left\| f(x_i) - f(x_j) \right\|_2$$

The online triplet loss encourages the distances between intra-class feature embedding to be smaller than that between the inter-class by at least $\alpha$.

It's well known that the sampling has important influence when training the triplet loss. Nevertheless, the exploration of mining hard examples is beyond this paper's scope. Hence, the lifted structured loss proposed in (Oh Song et al. 2016) is employed to ease the sampling problem. Our final optimization goal for structured embedding learning can be represented as:

$$L_{structured} = \frac{1}{2\left|\mathcal{P}\right|} \sum_{(i,j)\in\mathcal{P}} \max\left(R_{ij} + D_{ij}, 0\right)^2 \quad (2)$$

where,

$$R_{ij} = \log\{ \sum_{(i,k)\in\mathcal{N}} \exp(\alpha - D_{ik}) + \sum_{(j,l)\in\mathcal{N}} \exp(\alpha - D_{jl}) \}$$

$(i,j) \in \mathcal{P}$ means $y_i = y_j$ and $(i,k) \in \mathcal{N}$ means $y_i \neq y_k$. $R_{ij}$ actually finds the hardest negative examples in the batch for $x_i$ and $x_j$. The whole framework is trained by optimizing structured loss fuction and softmax loss function.

$$L = \lambda_1 L_{structured} + \lambda_2 L_{soft\max} \quad (3)$$

In practice, $\lambda_1$ and $\lambda_2$ are both set as 1.

By additionally learning structured embedding, the features are not only required to classify a scene correctly but also forced to maintain intra-class and inter-class similarities. Hence, the learned features are capable of discovering and representing more discriminative areas in the scene. To verify the effect, we alter the framework in Figure 3. The fully connected layers are replaced by fully convolutional layers and global average pooling is carried out to get the

[1] we use HHA (Horizontal Disparity, Height above the ground, Angle of surface norm) (Gupta et al. 2014) to encode the depth images in this paper.

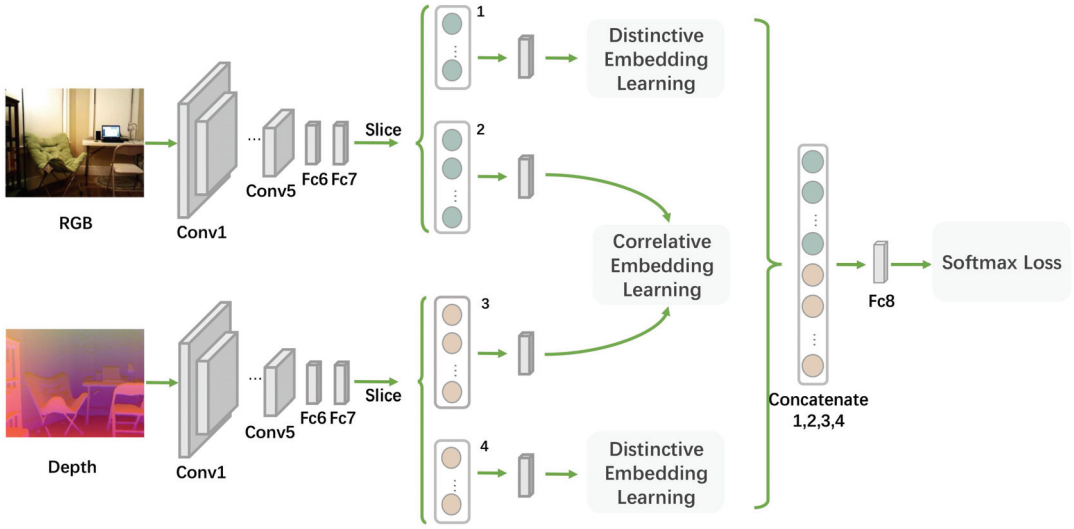


Figure 4: The proposed discriminative feature fusion network. The features from fc7 layer are divided into two parts to learn the distinctive embedding and correlative embedding respectively.

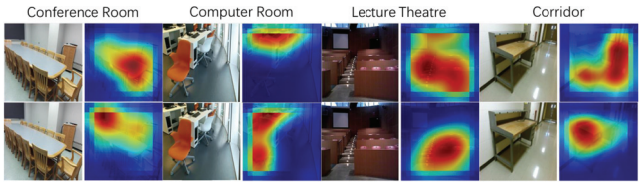


Figure 5: The class-specific activation map (CAM) is obtained from the modified network trained with (the top row) and without (the bottom row) structured embedding learning stream. The colors covered in images illustrates the degrees of importance of the area for classification. It can be observed that the structured embedding help to mine more discriminative areas to represent scenes.

final classification score. In the training phase, the proposed structured embedding learning stream is connected to the modified corresponding fully convolutional layers. To visualize which areas in the scene play more role in representation, the Class-specific Activation Map (CAM) proposed in (Zhou et al. 2016) is employed. The visualization is shown in Figure 5. The colors covered in images show the degrees of importance of the area for classification. It can be observed that without learning structured embedding, the features highlight on the frequently appeared and salient objects in the scene (such as *rows of chairs* in *lecture theatre*, *chair* and *computers* in *computer room*). Those parts are representative for the corresponding scenes, yet, they are not discriminative to discern the scene from other scenes. For example, *classroom* also has *rows of chairs*, and a lot of indoor scenes have *chair*. After structured embedding learning, however, more discriminative parts are mined. As illustrated in Figure 5, it can automatically locate *stairs* for *lecture theatre*, and ignore the noisy part *chair* for *computer room*.

## Multi-modal Discriminative Feature Fusion

After learning discriminative representation for each modality of RGB and Depth, directly concatenating them can already obtain more discriminative representation for RGB-D. However, the separately learned RGB and Depth features ignore the correlations between them. To explore the correlations between the RGB and Depth while simultaneously retain the distinctiveness in the two modalities, we propose a discriminative fusion network in the second learning stage, as illustrated in figure 4. The network consists of two streams which are combined in the fc7 layer. The weights in RGB and Depth streams are initialized with the trained RGB and Depth network in the first stage respectively. The fc7 layers of both RGB and Depth streams are divided into two parts. One of the two parts is supposed to represent the distinctive information in each modality, and the other part is designed to represent correlative information between multi-modalities.

**Distinctive Embedding Learning** To represent the distinctive information rather than the noisy information in the distinctive part, a similar pipeline as the first learning stage is proposed. The fully connected layer is connected to the distinctive part to reduce the dimension. The distinctive embedding then is learned under the supervision of the structured loss which has the same formula with function (2). We denote the loss as $L_{rgb\_dis}$ and $L_{d\_dis}$ for RGB and Depth streams respectively.

**Correlative Embedding Learning** Ideally, the multiple modalities of the same class ought to have strong correlation and the modalities of different classes should be less correlative. Denote the RGB and Depth input as random variables $\mathbf{x}$ and $\mathbf{z}$, the feature embedding of them are $f_{W_x}(\mathbf{x})$ and $f_{W_z}(\mathbf{z})$. According to Canonical Correlation Analysis (CCA), when sampling instances $(\langle x_1, z_1\rangle, \langle x_2, z_2\rangle, ..., \langle x_s, z_s\rangle)$ from the distribution of $\mathbf{x}$

and $\mathbf{z}$, we obtain the embedding for the instances as $f_{W_x}(\mathbf{x_S})$ and $f_{W_z}(\mathbf{z_S})$. The correlation of $\mathbf{x}$ and $\mathbf{z}$ can be formulated as:

$$\rho = \frac{\langle f_{W_x}(\mathbf{x_S}), f_{W_z}(\mathbf{z_S}) \rangle}{\|f_{W_x}(\mathbf{x_S})\| \, \|f_{W_z}(\mathbf{z_S})\|} \tag{4}$$

For $\mathbf{x}$ and $\mathbf{z}$ from the same class, $\rho$ is supposed to be maximized, or vice versa. To maximize the $\rho$, it actually requires the $f_{W_x}(\mathbf{x_S})$ and $f_{W_z}(\mathbf{z_S})$ to be collinear, thus the sampled instances in $\mathbf{x}$ and $\mathbf{z}$ should be as close as possible. On the contrary, sampled instances in $\mathbf{x}$ and $\mathbf{z}$ ought to be distant from each other.

Based on this idea, we propose the optimization goal for correlative embedding learning which can be formulated as:

$$\begin{aligned} L_{corr} = \frac{1}{4\,|\mathcal{P}|} \sum_{(x_i,z_j)\in\mathcal{P}} \{ & \max\left(R_{x_i,z_j} + D_{x_i,z_j}, 0\right)^2 \\ & + \max\left(T_{x_i,z_j} + D_{x_i,z_j}, 0\right)^2 \} \end{aligned} \tag{5}$$

where,

$$\begin{aligned} R_{x_i,z_j} &= \log\{ \sum_{(x_i,z_k)\in\mathcal{N}} \exp(\alpha - D_{x_i,z_k}) \} \\ T_{x_i,z_j} &= \log\{ \sum_{(x_m,z_j)\in\mathcal{N}} \exp(\alpha - D_{x_m,z_j}) \} \\ D_{x_i,z_j} &= \|f_{W_x}(x_i) - f_{W_z}(z_j)\|_2 \end{aligned}$$

The $R_{x_i,z_j}$ and $T_{x_i,z_j}$ are used to mine hard negative examples from Depth and RGB modalities. The loss for correlative embedding learning encourages the distance of intra-class RGB and Depth embedding to be closer by at least margin $\alpha$ than the inter-class RGB and Depth embedding. Notice that the correlation considered in (Wang et al. 2015b) and (Zhu, Weibel, and Lu 2016) only between RGB-D pairs, in function (5), $x_i$ and $z_j$ can be sampled randomly, as long as they belong to the same class. Thus, the correlation between $\mathbf{x}$ and $\mathbf{z}$ in the same class or different classes is better represented.

The final optimization goals for the second stage learning is composed of three parts: 1) the distinctive embedding learning for each modality supervised by the $L_{rgb\_dis}$ and $L_{d\_dis}$, 2) the correlative embedding learning between the multi-modalities supervised by the $L_{corr}$, 3) the classification learning supervised by $L_{softmax}$. They can be learned jointly as:

$$L = \beta_1 L_{rgb\_dis} + \beta_2 L_{d\_dis} + \beta_3 L_{corr} + \beta_4 L_{softmax} \tag{6}$$

In practice, the $\beta_1$, $\beta_2$, $\beta_3$, $\beta_4$ are all set as 1.

In the test phase, only the second stage framework is used, all the embedding learning streams are removed. It needs to be mentioned that although our framework is trained in two stages, all the parameters in the final model are optimized jointly.

# Experiments

The proposed method is validated on two popular RGB-D scene classification datasets: SUN RGB-D and NYU Depth Dataset V2.

## Datasets

The SUN RGB-D dataset contains 10,355 RGB and Depth image pairs captured from different cameras including Kinect v2, RealSense, Kinect v1 and Asus Xtion. We follow the experimental settings in (Song, Lichtenberg, and Xiao 2015). 19 categories are kept for our experiments with 4,845 images for training and 4,659 images for testing.

The NYU Depth Dataset V2 includes 1,449 RGB and Depth image pairs. To compare with other methods, we follow the experimental settings in (Gupta, Arbelaez, and Malik 2013). There are 795 training images and 654 testing images for 10 scene categories.

## Implementation Details

We implement the whole architecture in popular framework Caffe (Jia et al. 2014). The code will be available in https://github.com/liarba/scene_recognition. To fairly compare the results with the others, the AlexNet is used as the base architecture. The input image pairs are resized to $256 \times 256$ and randomly cropped into $227 \times 227$ as the input to the network. For SUN RGB-D Dataset, in the first stage, the weights are initialized using Places CNN and the batch size $n_0$, initial learning rate $\gamma_0$, stepsize $n_s$ and max iterations $N$ are respectively set as 256, 0.00001, 30000, 40000 for both RGB and Depth modality. In the second stage, the dimension ratio between the discriminative part and correlative part is set as 1:7. $n_0$, $\gamma_0$, $n_s$ and $N$ are set as 128, 0.0001, 3000, 4000. For NYU Depth Dataset V2, the $n_0$, $\gamma_0$, $n_s$ and $N$ are set as 256, 0.00001, 3000, 4000 in the first stage and they are modified as 128, 0.00001, 2000, 3000 in the second stage. The margin $\alpha$ is set as 1 according to the cross validation. Both the SUN RGB-D Dataset and NYU Depth Dataset V2 have highly imbalanced number of images between classes. To train the network effectively for the classes that have less number of images, we use frequency weighted softmax loss:

$$L_{softmax\_weighted} = \frac{1}{N} \sum_i -w(y_i) \log\left(\frac{f(x_i)_{y_i}}{\sum_j f(x_i)_j}\right) \tag{7}$$

where $w(t)$ is defined as:

$$w(t) = \frac{N_t - N_{c\_\min} + \delta}{N_{c\_\max} - N_{c\_\min}}$$

in which, $N_t$ is the number of images of class $t$ in the training set. $c\_\min$ or $c\_\max$ represents the class with the least or the most number of training images. The $\delta$ is set as 0.01. Following the previous work on RGB-D scene classification, we use the mean accuracy over categories as the evaluation metric.

## Results on SUN RGB-D dataset

*Comparison with State-of-the-art Methods* We compare our final results with the previous best results, as illustrated in Table 1. (Wang et al. 2015b) is originally proposed for RGB-D object recognition and re-implemented by (Zhu, Weibel, and Lu 2016) for RGB-D scene classification. Although

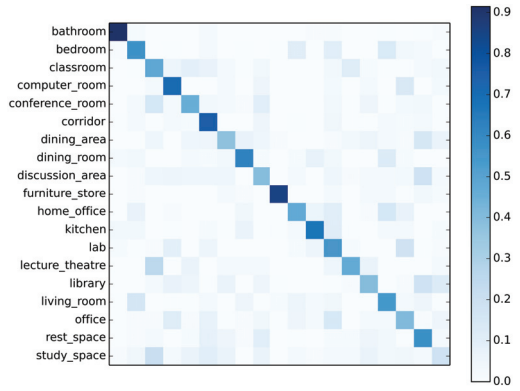


Figure 6: The classification confusion matrix of DF$^2$Net on SUN RGB-D Dataset. The vertical axis shows the ground-truth classes, the horizontal axis show the predicted classes.

(Wang et al. 2015b) gets high performance on RGB-D object recognition tasks, it fails to perform well on the more cluttered and complicated indoor scene. (Liao et al. 2016) uses a multi-task framework by taking the semantic segmentation of images as the regularizer for scene classification, which uses extra pixel-level annotation information. (Zhu, Weibel, and Lu 2016) improves the (Wang et al. 2015b) by adding the within-class and between-class correlations for image pairs. (Wang et al. 2016) uses the region proposal based CNN features to represent the scene while keeping the component sparsity and modal non-sparsity. Our proposed two-stage training framework further pushes the frontier and acquires 54.6% Mean Class Accuracy, outperforming all the state-of-the-arts by a significant margin. The confusion matrix can be seen in Figure 6. Besides superior performance, our method also consumes much less time in inference than the previous best performed method (Wang et al. 2016). For testing, our current net takes only 0.006s for each image. While (Wang et al. 2016) needs to extract proposals for each image, which is time consuming.

*Ablation Study* The ablation study results are shown in Table 2. For conciseness, we denote "Discriminative Feature Learning", "Correlative Feature Learning" and "Distinctive and Correlative Feature Leaning" as "DFL", "CFL" and "DCFL" for short respectively in the table.

The baselines for single modalities ("RGB AlexNet" and "Depth AlexNet") fine-tune the Places CNN with RGB and Depth images from SUN RGB-D dataset. They get 42.6% and 38.4% accuracy on SUN RGB-D. "DFL for RGB" and "DFL for Depth" applie our proposed structured embedding learning stream on the two modalities respectively. It can be seen in the Table 2 that for both modalities, DFL can improve the performance, quantitatively validating that our proposed discriminative feature learning network is able to learn more discriminative representation for each modality.

After obtaining the representation for RGB and Depth modalities. The multimodal representation learning strategies illustrated in Figure 2 are evaluated.

1) *"Baseline RGB+D"* In our experimental setting, the baseline method of multimodal fusion utilizes two-stream CNNs that fuse at the fc7 layer. The weights in each stream are initialized with the fine-tuned Places CNN on RGB and Depth modalities respectively. The concatenated fc7 layer then connects to fc8 layer to optimize the softmax loss function. As shown in Table 2, simply concatenating the features of two modalities can already get decent performance (48.3%), giving credit to the jointly optimizing in our model. Nevertheless, the learned representation is far from optimal.

Table 1: Performance comparison with the state-of-the-art methods on SUN RGB-D Dataset

| Methods | Accuracy(%) |
|---|---|
| (Wang et al. 2015b) | 26.5% |
| (Liao et al. 2016) | 41.3% |
| (Zhu, Weibel, and Lu 2016) | 41.5% |
| (Wang et al. 2016) | 48.1% |
| **DF$^2$Net (ours)** | **54.6**% |

Table 2: Ablation study on SUN RGB-D Dataset

| Methods | Accuracy(%) |
|---|---|
| RGB AlexNet | 42.6% |
| DFL for RGB | 46.3% |
| Depth AlexNet | 38.4% |
| DFL for Depth | 39.2% |
| Baseline RGB+D | 48.3% |
| DFL for RGB+D | 51.7% |
| DFL for RGB+D & DFL for RGBD | 51.7% |
| DFL for RGB+D & CFL for RGBD | 52.9% |
| DFL for RGB+D & DCFL for RGBD | **54.6**% |

2) *"DFL for RGB+D"* It implements Figure 2(c). Compared with baseline method, it initializes the RGB and Depth streams with representation learned from the proposed discriminative feature learning network. The Table 2 shows that "DFL for RGB+D" can already get better performance (51.7%). The performance outperforms the baseline by 3.4%, quantitatively verifying our observation that DFL for each modal will help to learn more discriminative representation for RGB-D.

3) *"DFL for RGB+D & DFL for RGBD"* Besides the methods analyzed in Figure 2, we also propose an alternative way to learn combined multimodal representation. The initialization of two-stream CNNs is the same as the 2). The proposed structured embedding stream then is also used after the concatenated features from fc7 layers. As shown in Table2, the DFL for combined feature has no effect. This might be caused by the DFL for single modality has already optimized the structured loss function for the combined features.

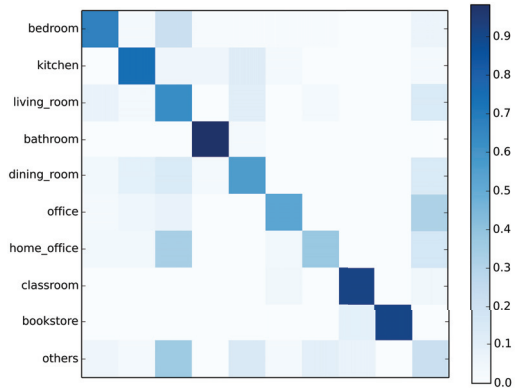


Figure 7: The classification confusion matrix of DF$^2$Net on NYU Depth Datasset V2.

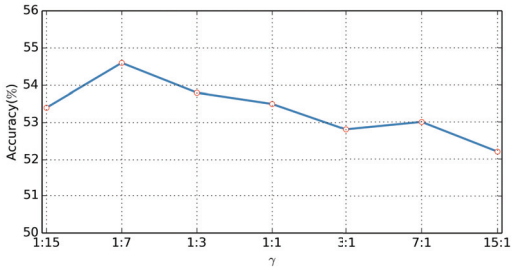


Figure 8: The effect of setting different ratios $\gamma$

4) *"DFL for RGB+D & CFL for RGBD "* It implements the Figure 2(d). In the implementation, after the same initialization of two-stream CNNs as the 2), the CFL is carried out to the holistic RGB and Depth representation by learning the proposed correlative embedding for them. The results in Table 2 show the performance can be improved slightly (1.2%) by considering correlation. This is because that some noises in each modality are able to be removed in representation. However, enforcing the RGB and Depth representation to be wholly consistent also impede to further learn more discriminative multimodal representation for scene classification.

5) *"DFL for RGB+D & DCFL for RGBD "* It represents our final modal. Compared with the structure in 4), we carry out DCFL for RGB and Depth fusion. By allowing representation to learn correlative part and distinctive part separately, we gets further improvement (1.7%) on the classification accuracy.

Our final performance gets 54.6% accuracy on SUN RGB-D Dataset, outperforming the baseline model with a large margin (6.3%).

*Hyper-parameter Analysis* The ratio $\gamma$ of distinctive part's dimension to the correlative part's dimension has great impact on the final results. When the ratio becomes infinite max, it actually equals to the methods represented in Figure 2(c) and it represents to learn no correlations between RGB and Depth. When the ratio decreases to 0, it actually acts in the the strategy shown in Figure 2(d). The classification accuracies with varying $\gamma$ are shown in Figure 8.

Table 3: Performance comparison with the state-of-the-art methods on NYU Depth Dataset V2

| Method | Accuracy(%) |
|---|---|
| (Gupta, Arbelaez, and Malik 2013) | 45.4% |
| (Wang et al. 2016) | 63.9% |
| **ours** | **65.4**% |

Table 4: Ablation study on NYU Depth Dataset V2

| Model | Accuracy(%) |
|---|---|
| RGB AlexNet | 59.5% |
| DFL for RGB | 61.1% |
| Depth AlexNet | 49.3% |
| DFL for Depth | 54.8% |
| Baseline for RGB+D | 60.6% |
| DFL RGB+D | 63.5% |
| DFL for RGB+D & DCFL for RGBD | **65.4**% |

### Results on NYU Depth Dataset V2

Table 3 shows the performance comparison with state of the arts on NYU Depth Dataset V2. Our model gets 65.4% mean class accuracy, also outperforming all the state of the arts. Notice that less margins over state of the arts are obtained on NYU Depth Dataset V2. It might be due to that there are too less training images and the "other" category contains less useful structured information. We also visualize the confusion matrix for the final results on NYU Depth Dataset V2 in Figure7. Table 4 shows the ablation study results on NYU Depth Dataset V2. From the results, we can get consistent conclusions with that of the SUN RGB-D Dataset.

## Conclusions

In this paper, we propose a Discriminative Feature Learning and Fusion Network (DF$^2$Net) with two-stage learning for the RGB-D indoor scene classification. In the first stage, besides optimizing the canonical softmax loss function, the structured embedding is also proposed. In the second stage, we obtain more discriminative multimodal features by learning the distinctive embedding and correlative embedding. We achieve the state-of-the-art results for RGB-D indoor scene classification on both SUN RGB-D Dataset and NYU Depth Dataset V2, which validates the effectiveness of the proposed framework.

## Acknowledgment

This work is funded by the National Natural Science Foundation of China (Grant No. 61403387), the National Key Research and Development Program of China (Grant No. 2016YFB1001004) and the Projects of Chinese Academy of Science (Grant No. QYZDB-SSW-JSC006, Grant No. 173211KYSB20160008).

## References


Arandjelovic, R.; Gronat, P.; Torii, A.; Pajdla, T.; and Sivic, J. 2016. Netvlad: Cnn architecture for weakly supervised place recognition. In *CVPR*, 5297–5307.

Bappy, J. H.; Paul, S.; and Roy-Chowdhury, A. K. 2016. Online adaptation for joint scene and object classification. In *ECCV*, 227–243.

Cheng, Y.; Cai, R.; Zhang, C.; Li, Z.; Zhao, X.; Huang, K.; and Rui, Y. 2015a. Query adaptive similarity measure for rgb-d object recognition. In *ICCV*, 145–153.

Cheng, Y.; Cai, R.; Zhao, X.; and Huang, K. 2015b. Convolutional fisher kernels for rgb-d object recognition. In *3DV*, 135–143. IEEE.

Cheng, Y.; Zhao, X.; Cai, R.; Li, Z.; Huang, K.; and Rui, Y. 2016. Semi-supervised multimodal deep learning for rgb-d object recognition. In *IJCAI*, 3345–3351.

Cheng, Y.; Cai, R.; Li, Z.; Zhao, X.; and Huang, K. 2017. Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation. In *CVPR*, 3029–3037.

Couprie, C.; Farabet, C.; Najman, L.; and LeCun, Y. 2013. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*.

Cui, Y.; Zhou, F.; Lin, Y.; and Belongie, S. 2016. Fine-grained categorization and dataset bootstrapping using deep metric learning with humans in the loop. In *CVPR*, 1153–1162.

Dixit, M.; Chen, S.; Gao, D.; Rasiwasia, N.; and Vasconcelos, N. 2015. Scene classification with semantic fisher vectors. In *CVPR*, 2974–2983.

Eitel, A.; Springenberg, J. T.; Spinello, L.; Riedmiller, M.; and Burgard, W. 2015. Multimodal deep learning for robust rgb-d object recognition. In *IROS*, 681–687.

Gordo, A.; Almazán, J.; Revaud, J.; and Larlus, D. 2016. Deep image retrieval: Learning global representations for image search. In *ECCV*, 241–257.

Gupta, S.; Arbelaez, P.; and Malik, J. 2013. Perceptual organization and recognition of indoor scenes from rgb-d images. In *CVPR*, 564–571.

Gupta, S.; Girshick, R.; Arbeláez, P.; and Malik, J. 2014. Learning rich features from rgb-d images for object detection and segmentation. In *ECCV*, 345–360.

Hadsell, R.; Chopra, S.; and LeCun, Y. 2006. Dimensionality reduction by learning an invariant mapping. In *CVPR*, volume 2, 1735–1742.

Huang, F.; Cheng, Y.; Jin, C.; Zhang, Y.; and Zhang, T. 2017. Deep multimodal embedding model for fine-grained sketch-based image retrieval. In *ACM SIGIR*, 929–932.

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

Li, L.-J.; Su, H.; Lim, Y.; and Li, F.-F. 2010. Objects as attributes for scene classification. In *ECCV Workshops*, 57–69.

Liao, Y.; Kodagoda, S.; Wang, Y.; Shi, L.; and Liu, Y. 2016. Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks. In *ICRA*, 2318–2325. IEEE.

Oh Song, H.; Xiang, Y.; Jegelka, S.; and Savarese, S. 2016. Deep metric learning via lifted structured feature embedding. In *CVPR*, 4004–4012.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; et al. 2015. Deep face recognition. In *BMVC*, volume 1, 6.

Schroff, F.; Kalenichenko, D.; and Philbin, J. 2015. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 815–823.

Song, S.; Lichtenberg, S. P.; and Xiao, J. 2015. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 567–576.

Wang, J.; Song, Y.; Leung, T.; Rosenberg, C.; Wang, J.; Philbin, J.; Chen, B.; and Wu, Y. 2014. Learning fine-grained image similarity with deep ranking. In *CVPR*, 1386–1393.

Wang, A.; Cai, J.; Lu, J.; and Cham, T.-J. 2015a. Mmss: Multi-modal sharable and specific feature learning for rgb-d object recognition. In *ICCV*, 1125–1133.

Wang, A.; Lu, J.; Cai, J.; Cham, T.-J.; and Wang, G. 2015b. Large-margin multi-modal deep learning for rgb-d object recognition. *IEEE Transactions on Multimedia* 17(11):1887–1898.

Wang, A.; Cai, J.; Lu, J.; and Cham, T.-J. 2016. Modality and component aware feature fusion for rgb-d scene classification. In *CVPR*, 5995–6004.

Wang, Z.; Wang, L.; Wang, Y.; Zhang, B.; and Qiao, Y. 2017. Weakly supervised patchnets: Describing and aggregating local patches for scene recognition. *IEEE TIP* 26(4):2028–2041.

Wu, R.; Wang, B.; Wang, W.; and Yu, Y. 2015. Harvesting discriminative meta objects with deep cnn features for scene classification. In *ICCV*, 1287–1295.

Yu, J.; Yang, X.; Gao, F.; and Tao, D. 2016. Deep multimodal distance metric learning using click constraints for image ranking. *IEEE transactions on cybernetics*.

Zhang, X.; Zhou, F.; Lin, Y.; and Zhang, S. 2016. Embedding label structures for fine-grained feature representation. In *CVPR*, 1114–1123.

Zhou, B.; Lapedriza, A.; Xiao, J.; Torralba, A.; and Oliva, A. 2014. Learning deep features for scene recognition using places database. In *NIPS*, 487–495.

Zhou, B.; Khosla, A.; Lapedriza, A.; Oliva, A.; and Torralba, A. 2016. Learning deep features for discriminative localization. In *CVPR*, 2921–2929.

Zhu, H.; Weibel, J.-B.; and Lu, S. 2016. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In *CVPR*, 2969–2976.