Aberrance-aware Gradient-sensitive Attentions for Scene Recognition with RGB-D Video

ID: 391

ABSTRACT

With the developments of deep learning, previous approaches have made successes in scene recognition with massive RGB data obtained from the ideal environments. However, scene recognition in real world may face various types of aberrant conditions caused by different unavoidable factors, such as the lighting variance of the environments and the limitations of cameras, which may damage the performance of previous models. In addition to ideal conditions, our motivation is to investigate researches on robust scene recognition models for unconstrained environments. In this paper, we propose an aberrance-aware framework for RGB-D scene recognition, where several types of attentions, such as temporal, spatial and modal attentions are integrated to spatio-temporal RGB-D CNN models to avoid the interference of RGB frame blurring, depth missing, and light variance. All the attentions are homogeneously obtained by projecting the gradient-sensitive maps of visual data into corresponding spaces. Particularly, the gradient maps are captured with the convolutional operations with the typically designed kernels, which can be seamlessly integrated into end-to-end CNN training. The experiments under different challenging conditions demonstrate the effectiveness of the proposed method.

KEYWORDS

Scene recognition, RGB-D video, attention, gradient-sensitive, challenging conditions

ACM Reference Format:

. 2019. Aberrance-aware Gradient-sensitive Attentions for Scene Recognition with RGB-D Video: ID: 391. In *Proceedings of ACM Multimedia conference (Conference'19).* ACM, New York, NY, USA, Article 4, 9 pages. https://doi.org/****

1 INTRODUCTION

Scene recognition is to annotate scene labels for visual data such as images or videos. Humans have talent to recognize coarse-grained scenes rapidly in a glance without hard training [5, 17, 29, 34], also can distinguish more fine-grained scenes in high accuracy [34]. In contrast, scene recognition is a challenging task for artificial vision systems, which usually requires massive data to train complex models (such as deep CNN models [12, 15, 23]). Although extracting features with CNN models pretrained on large-scale database (such as Places [35]) has been shown to be powerful on many other scene recognition tasks, such as 15 scenes [16], MIT67 [19] and SUN397 [32], current works only investigate researches and evaluate the obtained models under the assumptions of ideal environments. In contrast to ideal environments, the aberrant conditions are unavoidable in real life, which may lead to great challenges to the scene recognition models.

Office (a) bright (b) dim — Conference room (a) bright (d) dark

Figure 1: RGB-D data of scene categories Office and Conference room captured from different lighting conditions. Frames in top/bottom rows are captured with RGB/Depth cameras in bright, dim and dark conditions.

RGB cameras are usually sensitive to the lighting variance of the environments. Even capturing the same scenes with different lighting conditions leads to large discrepancy in RGB data. Previous works that focus on RGB data may be easily interfered by light variance due to the lighting sensitiveness of RGB cameras. In contrast, depth cameras such as Microsoft Kinect, consisting of an infrared laser projector combined with a monochrome Complementary Metal Oxide Semiconductor (CMOS) sensor, can capture more consistent visual data in different ambient-lighting conditions. Thus integrating depth with RGB data is typically helpful to recognize scenes in lighting variance environments. However, the development of RGB-D recognition is not as successful as RGB recognition due to some known limitations, such as lack of depth data, short response range of depth camera, and some other aberrant conditions caused by the factors such as lighting variance of the environments or hardware limitation of cameras.

Due to the lack of depth data, earlier works [11, 25, 31, 36] train depth CNN models through fine tuning or transferring the pretrained RGB models to depth. Recent work [26] argues that fine tuning or transferring are not capable of learning depth modality-specific features, then a more shallow depth-CNN is proposed, which is particularly trained from scratch with data augmentation in weak supervision. With the depth-CNN trained from scratch,

Conference'19, October 2019, Nice, France

ACM ISBN 123-4567-24-567/08/06…$15.00

https://doi.org/****

later work [27] proposes to combine multi-source multi-modal CNNs with skip connections. In addition to data limitation, a more recent work [28] argues that the short response range (usually less than 5m) of depth camera may damage the performances of wide scene recognition. And they propose a framework to capture wide scenes with videos by traveling across the camera, where the videos are encoded with a joint CNN-LSTM architecture for scene recognition. These known limitations have been somewhat addressed in above works. However, these previous works still investigate researches on scene recognition models in ideal conditions, lacking the researches and adaptations to the unavoidablely and dynamically aberrant conditions in real world.

Typically, the aberrant conditions usually lead to large appearance (pixel) changes in the visual source data (images or videos). For instance, the lighting variance leads to significant differences in RGB data (but not depth data,see Fig. 1); the traveling of cameras (during video capturing) may lead to RGB frame blurring due to the speed mismatch between lens focusing and camera moving; also the traveling of cameras may lead to regional depth vanishing due to limited response range of depth camera (note that although wide scenes can be completely captured with the whole videos [28] in final, while the regional depth missing during camera moving may easily happen). However, recent CNN based RGB-D scene recognition works only focus on feature learning and fusion, where the appearance changes (due to the aberrant conditions) may not be vanished in the hidden features of CNNs. Thus the visual appearance changes can not be sensed in those previous works, limiting the performances of scene recognition in aberrant condition.

In contrast, our motivation is to dynamically sense the aberrant conditions, and investigate research on more robust recognizing models for different environments. Particularly, one observation is that different types of aberrant conditions usually result in the gradient changes between pixels. For instance, RGB frame blurring results in frame-level gradient decrements, depth vanishing results in region-level gradient missing, and lighting variance results in gradient increment (with brighter light) or decrement (with darker light). Although those aberrant conditions are not the only necessary factors of gradient changes, specific aberrant condition still can be sensed by measuring gradients in corresponding dimensions of RGB-D video data.

In this paper, we propose an aberrance-aware framework for scene recognition with RGB-D videos. First, several particular operators are implemented to obtain gradient-sensitive maps of video frames, which are designed as auxiliary convolutional kernels, allowing us train the proposed model end-to-end. Then, the gradient-sensitive maps are projected into different dimensions to obtain several types of attentions, such as temporal, spatial and modal attentions, which are then integrated to our RGB-D spatio-temporal CNN architecture to avoid the interference of aberrant conditions, such as RGB frame blurring, depth vanishing, and light variance. Due to the data redundancy, the proposed attending manner is particularly beneficial to the research of video based scene recognition, where weakening the damaged data (such as blurring frames, depth vanished regions and RGB frames in dark) meanwhile emphasizing effective data is more powerful and efficient, thus attempts to fix the damage data with external algorithms are not necessary. For the evaluations, we collect RGB-D video data in different challenging conditions. In addition to collected data, we also make attempts to generate synthetic data to simulate different challenging conditions, which are integrated to train the proposed model. And the experimental results demonstrate that the proposed model trained with those synthetic data is also effective to the "real" data collected from different aberrant conditions in test. The contributions of this paper can be summarized as follows:

- In contrast to previous works that focus on ideal environments, we propose an aberrance-aware framework for RGB-D scene recognition, which is more robust to different aberrant conditions.
- We propose several gradient-sensitive attentions to avoid the interference of aberrant conditions, such as frame blurring, depth vanishing or lighting variance.
- We demonstrate that training the proposed model with the merging of ideal and synthetic data is also effective to the test of "real" data collected from challenging conditions.

2 RELATED WORKS

2.1 RGB-D scene recognition

Comparing to RGB data, depth data is rarely affected by the lighting variance, which typically benefits the recognition of indoor scenes with significant or frequent changes of illumination. Some works about RGB-D data [1, 10] focus on handcrafted features, which are engineered specifically for different tasks by experts to capture some specific properties. With the success of deep learning models in RGB data, more recent RGB-D approaches have investigated researches on deep learning models, particularly on CNN models. Compared to RGB data, the lack of training data limit the performance of RGB-D scene recognition.

Socher *et al.* [24] propose an architecture with one single convolutional layer, which is unsupervisedly trained with patches, and then fused with a recurrent neural network (RNN). Later, with the introduction of SUN RGB-D database [25], many recent works [11, 26, 27, 31, 36] have made attempts to train deeper CNN models for RGB-D scene recognition. Some works [11, 25, 31, 36] train depth CNN models by fine tuning from pretrained RGB models to depth. With the fine tuned CNN models, one approach [31] extracts deep features on both local regions and whole images on both RGB, depth and surface normal, which are further combined with component-aware fusion. Some approaches [11, 36] propose to incorporate CNN architectures to jointly fine tune RGB and depth image pairs. Zhu *et al.* [36] jointly fine tune the RGB and depth CNN models with a multi-modal fusion layer, which models both inter and intra-modality correlations, and results in more compact and discriminative features. Alternatively, Gupta *et al.* [11] propose a cross-modal distillation approach where depth model learning is guided by the high-level RGB features obtained from the paired RGB image. One recent work [26] argues that fine tuning the pretrained RGB model with depth data may not obtain depth modality-specific features. In contrast to fine tuning, the authors propose to train depth CNN models from scratch with patches in weak supervision. Song *et al.* [27] propose to train CNN with Alex-CNN [15] architecture, which combines with the fine tuned Places-CNN as a multi-source CNN model.

Then a more recent work [28] argues that another limitation of common RGB-D cameras is the limited response range. Particular, the authors propose to capture wide scenes with videos, where each frame only captures part of scenes, and more complete scenes are accumulated with the camera traveling. They also propose a multi-modal CNN-RNN architecture for RGB-D video recognition and introduce a new RGB-D video benchmark for scene recognition. Although some earlier RGB-D datasets [21, 33] also contain video data, which mainly focus on segmentation tasks, and are in lack of category diversity.

Although implementing RGB-D data for scene recognition, above works still lack of researching for different aberrant conditions. Since only focusing on hidden feature extraction and fusion, the key information reflected on visual appearance are ignored. In contrast, we investigate researches for appearance (gradient) sensitive attention manners to jointly address the problem of lighting variance, frame blurring and depth vanishing.

2.2 Video recognition

2.2.1 Video dataset. Several RGB video datasets for scene recognition have been introduced in previous works [3, 7, 20]. Moving vistas [20] focuses on scenes with highly dynamic patterns, such as fire, crowded highways or waterfalls, using chaos theory to capture dynamic attributes. Derpanis *et al.* [3] study how appearance and temporal dynamics contribute to scene recognition. Feichtenhofer *et al.* [7] propose a new dataset with more categories and an architecture using residual units and convolutions across time. However, RGB data of these datasets are captured in ideal environments, which are not capable of supporting the research of scene recognition under aberrant conditions, such as lighting variance. Although a RGB-D video database is introduced for scene recognition, the data is still captured under bright conditions, which cannot completely satisfy the requirements of our research. In this paper, we particularly capture new RGB-D video data under different aberrant conditions for different evaluations.

2.2.2 Spatio-temporal embedding. An intuitive solution of video recognition is to embed spatio-temproal information to predict labels. The spatio-temproal information can be categorized into different aspects. Some earlier works of action recognition [4, 14, 22] separately encode frames with static CNN models to extract frame-wise features. And also frames are extended to video clips, which are considered as multi-channel inputs, to feed 2D CNN models. Action recognition is an example that requires modeling appearance and temporal dynamics. Two streams base architecture is also proposed to jointly embed appearance and motion [8]. Later, Feichtenhofer *et al.* [7] propose to extend the two stream CNN with a ResNet architecture [12] and apply it to scene recognition. Also, the temporal information can be modeled with sequential models [18, 28], such as recurrent neural network (RNN). Those approaches focus on 2D convolutions, where the temporal information is externally encoded.

In contrast, feed-forward models are operated with 3D convolutions (C3D) [13, 30] in width, height, and temporal dimensions, although the 3D convolutional kernels may also be "inflated" from pretrained 2D kernels [2, 6].

Table 1: 3D CNN architecture.

Layer	Kernel size	Stride	Output size
Input	-		N×3×16×112×112
Conv 1:1	3×3×3	[1,1,1]	N×64×16×112×112
Pool 1	1×2×2	[1,2,2]	N×64×16×56×56
Conv 2:1	3×3×3	[1,1,1]	N×128×16×56×56
Pool 2	2×2×2	[2,2,2]	N×128×8×28×28
Conv 3:1-2	3×3×3	[1,1,1]	N×256×8×28×28
Pool 3	2×2×2	[2,2,2]	N×256×4×14×14
Conv 4:1-2	3×3×3	[1,1,1]	N×512×4×14×14
Pool 4	2×2×2	[2,2,2]	N×512×2×7×7
Conv 5:1-2	3×3×3	[1,1,1]	N×512×2×7×7
Pool 5	2×2×2	[2,2,2]	N×512×1×4×4

In 2D CNN based works, each frame is first encoded to a global feature, and then feed to the temporal embedding models. Separately emphasizing temporal information may be more effective to action recognition since the sequences are also key to action recognition. However, in our goal of scene recognition with videos is to extend the scope of depth cameras, where more frames of videos are particularly captured to obtain more wide scenes. It seems that explicitly temporal embedding is not as important as more completely spatial embedding, thus, we implement 3D convolutions on our RGB-D videos to simultaneously embed spatio-temporal information.

3 3D RGB-D SCENE RECOGNITION ARCHITECTURE

Previous work [28] has made attempt on RGB-D video recognition with a CNN-LSTM framework, where CNN is to embed spatial information with multiple convolutional layers and LSTM is to embed temporal information with recurrent modules. However, in CNN-LSTM frameworks, the temporal information is only embedded with the global CNN features, where temporal information in regions is ignored. Since our goal of recognizing scenes with video data is to extend scope of depth camera, temporal embedding in regions is also required. Thus instead of CNN-LSTM framework, we implement our RGB-D video recognition framework based on 3D CNN architecture [30].

Preliminary architecture. According to the procedure of recording a RGB-D video, the scene is scanned by moving the camera to get richer scene information. In order to extract more abundant scene information, an architecture is needed to extract temporal feature from the video effectively. In contrast to CNN-LSTM framework in [28], our goal is to jointly embed spatio-temporal information, since the temporal information is the extension of spatial information. Separately embedding temporal information is not as necessary as jointly embedding it with spatial information. Therefore, conventional 2D convolutional kernels and pooling layers cannot satisfy our goal. Inspired by [30], an architecture consisting of 3D convolutional kernels and pooling layers is implemented in our work.

Particularly, the preliminary architecture of 3D CNN is illustrated in Table 1. With the input of video clips in size of $N \times C \times L \times H \times W$

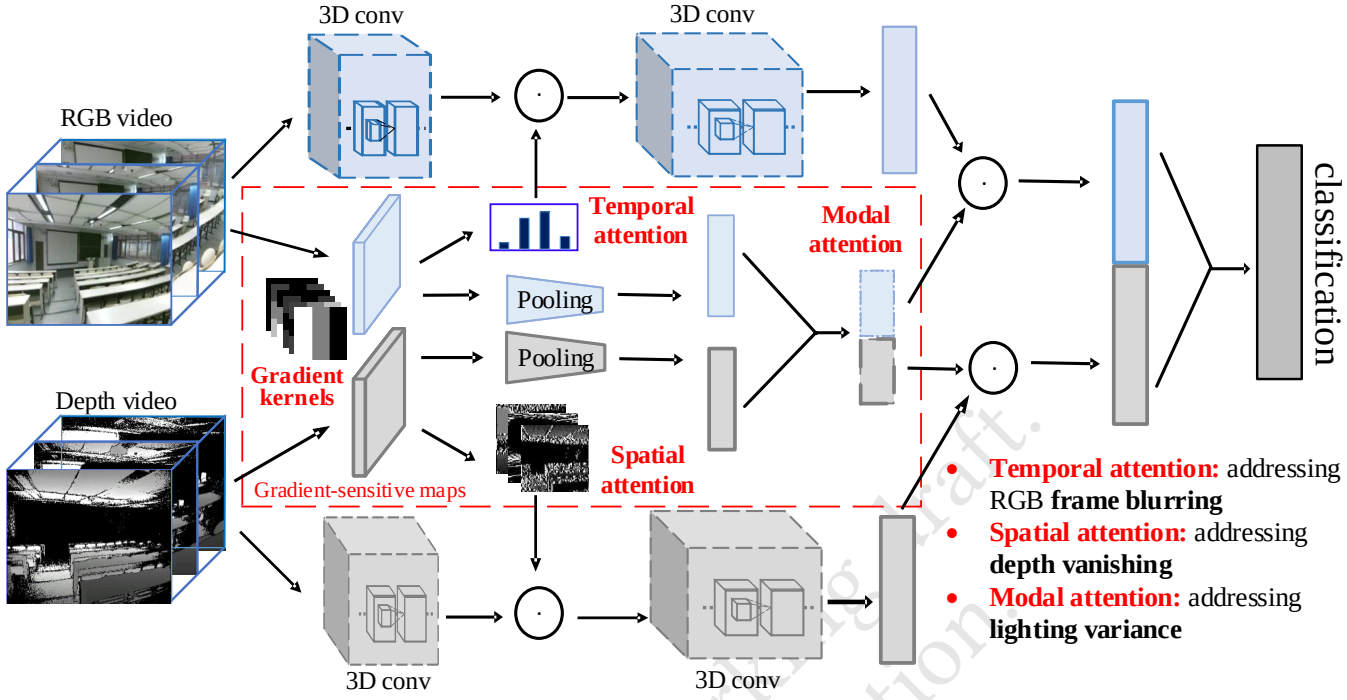


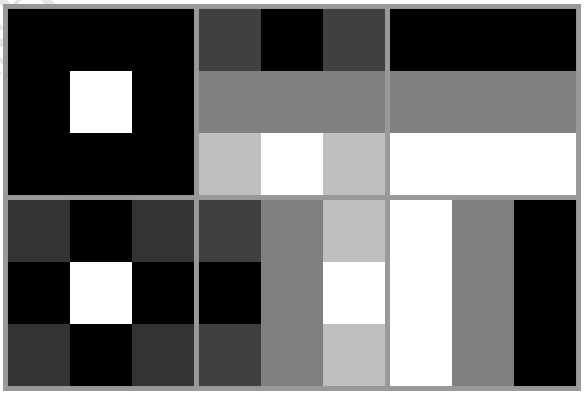
Figure 2: The proposed framework with gradient-sensitive temporal, spatial and modal attentions to address the aberrant conditions of frame blurring, depth vanishing and lighting variance.

($N = 8$ is the batch size, $C = 3$ is the number of channels, note that depth data is also transformed into three dimensions with fake colors, $L = 16$ is the number of frames of the video clips, and H, W are the height and width respectively), five convolutional layers and five pooling layers are implemented for video embedding, two fully connected layers and one softmax layer are implemented to predict scene labels. Kernel size ($C \times H \times W$), stride ($C \times H \times W$) and output size ($N \times C \times L \times H \times W$) of different layers are also illustrated. The stride of first pooling layer is particularly defined as $[1, 2, 2]$ to completely obtain frame-wise output, which is further adapted with attentions (will be introduced in next section).

RGB-D multi-modal 3D architecture. In our case, the preliminary 3D architecture is extended to RGB-D multi-streamed architecture. In particular, training our baseline RGB-D 3D architecture consists of two steps: 1) separately training modality-specific stream; 2) concatenating the fully connected layer of each stream for joint training (fine tuning). .

4 ABERRANT-AWARE ATTENTIONS WITH GRADIENT-SENSITIVE KENERLS

The aberrant conditions such as lighting variance, frame blurring and depth vanishing can damage the video data, limiting the performance of RGB-D scene recognition in real world. Since the continuous frames of scene videos usually contains data redundancy, emphasizing available data (while downplaying damaged data) seems more efficient than directly fixing damaged data. In this section, we propose several attentions to sense aberrant conditions and avoid the interference caused by them.



(a) Laplacian (b) Roberts (c) Sobel

Figure 3: Visualizations of the convolutional kernels.

4.1 Gradient-sensitive kernels

We typically design convolutional kernels $K = [k_1, k_2, ...k_6]$ based on several types of hand-crafted operators, such as Laplacian, Roberts and Sobel operators, which are visualized in Fig. 3. These kernels are implemented as auxiliary paths on visual data (video frames) to obtain the gradient maps $G = K \otimes X$ ($X \in R^{3\times16\times112\times112}$ is the split

video clips with 16 frames, $G \in R^{6\times16\times112\times112}$), which can somewhat measure the level of blurring, depth vanishing and brightness. Based on the obtained G, different types of attention are calculated by projecting G into different dimensions.

4.2 Temporal attention for frame blurring

Since we have to remove RGB-D cameras to capture the wide scenes, the frame blurring is unavoidable. The faster we move the cameras, the easier we obtain the blurring frames. Although slowing down the moving speed can somewhat alleviate such problem, it also leads to much longer process time to recognize scenes. Thus, we propose temporal attention to capture scene in a reasonable speed without affecting by blurring frames.

Since the scene information are redundantly stored in videos, the continuous frames usually contain large overlap. It's not necessary to directly enhance the blurring frames, just emphasizing the clear frames is more convenient. Emphasizing clear frames while downplaying blurring frames can be carried out with attention manner. Typically, frame blurring usually results in decrements of gradients, the gradient map G can somewhat measure the level of blurring of each frame. Thus we particularly propose temporal attention by projecting gradient map G into the dimension of frames.

The gradient map G of each video is first aggregated with average pooling in width, height and channel dimensions by:

$$t^i = \frac{\sum_{c,h,w\in R^{C\times H\times W}} G^{(c,i,h,w)}}{C\times H\times W}$$

where W, H, C are the numbers of width,height and channels, $T = \left[t^1, ...t^i, ..t^L\right]$ is aggregated gradient value, L is the number of frames in a short video. The global temporal attention is obtained by $A_t = softmax\,(T)$. Thus, for the output feature maps F^i_{rgb-1} (of conv1) of each frame, temporal attention is implemented with dot products with F^i_{rgb-1}, obtaining $\bar{F}^i_{rgb-1} = A^i_t \odot F^i_{rgb-1}$.

4.3 Spatial attention for depth vanishing

Although each frame particularly captures part of scenes, it's unavoidable to result in incomplete depth information during camera moving. The missing of depth is caused by the response range exceeded of depth camera, which leads to inaccurate distance estimation and damage the performance of recognizing models. Typically, the missing regions are usually represented with pixels of constant value (zero), also leading to very small gradients (almost none) in these regions. Similarly, we can also measure the level of depth missing with gradient map G. Thus we particularly propose spatial attention by projecting gradient map G into spatial related dimensions.

For each frame, the gradient map G is aggregated with average pooling in channels by $s^{(i,h,w)} = \frac{\sum_{c\in R^C} G^{(c,i,h,w)}}{C}$, then the spatial attention is obtained by $A^i_s = \begin{bmatrix} s^{(i,1,1)} & \dots & s^{(i,1,W)} \\ \vdots & s^{(h,w,i)} & \vdots \\ s^{(i,H,1)} & \dots & s^{(i,W,H)} \end{bmatrix}$.

Thus, for the output feature maps F^i_{conv1} (of conv1) of each frame, spatial attention is implemented with dot products with F^i_{d-1}, obtaining $\bar{F}^i_{d-1} = A^i_s \odot F^i_{d-1}$.

4.4 Modal attention for lighting variance

Typically, our goal is to emphasize the meaningful data while alleviating the interference of mess data. In addition to blurring and depth vanishing, the lighting variance also damages the performance, which is particularly sensitive to RGB data. Generally, all these matters can be measured with the gradients after projecting in different dimensions. Those aberrant conditions (e.g., lighting variance, blurring and depth vanishing) usually lead to the low gradients of the data. Since we synchronously obtain pairs of RGB and depth videos (or frames), we propose modal attention to emphasize the more useful modality in each time stamp. The modal attention is still designed with the gradient map G, and RGB-D feature fusion is implemented on the output of last convolutional layer. In order to match the size of output feature maps, we also aggregate the gradient map G with the average pooling, obtaining the follows:

$$m^{(h_0,w_0)} = \frac{\sum_{(i,h,w)\in R^{(W_{r5}\times H_{r5}\times L)}} G^{(i,h_0+h,w_0+w)}}{L\times H_5\times W_5}$$

where $H_{r5}\times W_{r5}$ is the receptive field of pool5 layer (i.e., 28×28), $M_{modality} = \begin{bmatrix} m^{(1,1)} & \dots & m^{(1,W_{o5})} \\ \vdots & m^{(h_0,w_0)} & \vdots \\ m^{(H_{o5},1)} & \dots & m^{(H_{o5},W_{o5})} \end{bmatrix}$. The modal attention $A_m = \left[A_{rgb}, A_d\right] = softmax\left(\left[M_{rgb}, M_d\right]\right)$ is implemented with dot products with $F_5 = \left[F_{rgb-5}, F_{d-5}\right]$, obtaining $\bar{F}_5 = A_m \odot F_5$.

5 EXPERIMENTAL RESULTS

5.1 Setting

5.1.1 Ideal data. In order to investigate scene recognition in RGB-D videos, a RGB-D video database is first introduced in [28]. That database consists of indoor videos captured from three different cities (separated up to 1000 km), guaranteeing diversity in locations and scenes. The database reuses 58 of the categories in the taxonomy of the MIT indoor scene database [19], and contains about 278 videos with more than five hours of footage in total. The duration of videos varies, depending on the complexity and extension of the scene itself (a *classroom* or *furniture store* requires more footage than *office* or *bedroom*) and how common and easy to access are certain categories (e.g., *office* and *classroom* have more videos than *auditorium* or *bowling alley*). Videos are captured using a Microsoft Kinect version 2 sensor, with a frame rate of 15 frames/s. Due to the long-tailed data distribution of that database (Eight scene categories contain only one video), only 50 categories are finally selected for model evaluation.

In order to conduct more convincing evaluations, we introduce a new RGB-D video database for scene recognition by collecting more data to different scene categories, where the scenes that contain very less data in previous database [28] and seven new scenes are particularly enhanced. Finally, we obtain 65 scenes with 423 videos for evaluation. In each category, about 70% of videos are randomly selected as training data, and the rest is considered as test data. Since most of data in [28] is captured under the bright environments, our complementary data is mainly captured in the similar environments to maintain the consistency.

- *Ideal-data*: for the rest of paper, above data is denoted as "ideal data".

5.1.2 Challenging data. Since our goal is to investigate research on scene recognition under the aberrant conditions, in addition to ideal data, we also capture more types of data in different aberrant conditions, such as capturing videos under the weaker lighting environments and with faster camera moving. Particularly, these complex data is also included to the evaluations, denoted as "challenging data" in rest of paper.

In addition to the challenging data captured in real world, we also make attempts to generate some synthetic data to simulate different lighting environments, which are quantized as three levels, such as bright, dim and dark. The synthetic data is transformed from the ideal data by implementing image processing operations to change the settings, such as brightness and contrast ratio. These data is denoted as "synthetic data" in rest of paper.

In summary, the challenging and synthetic data can be categorized into the following groups:

- *C-Fast-data*: captured with fast camera moving, with about 1.5x speed;
- *C-Faster-data*: captured with faster camera moving, with about 2.5x speed the faster moving the higher probability to obtain more blurring frames;
- *C-Dim-data*: captured under the dim condition;
- *C-Dark-data*: captured under the dark condition;
- *S-Dim-data*: synthetic data generated to simulate the dim condition;
- *S-Dark-data*: synthetic data generated to simulate the dark condition;

5.1.3 Preprocessing and evaluation metric. All the RGB-D videos are divided into clips consisting of 16 frames for 3D CNN training and evaluations. With 423 full videos, we obtain 12092 video clips, training/test split is 8822/3270. For the calibration of each pair of RGB-D frames, a pair of cropped strategies are implemented to the RGB and depth frames, respectively. The frames are first cropped into the size of 540×701, and are finally resized to 112×112. Inspired by previous RGB-D works [9, 28], we also extend our depth data to three channels. Since HHA encoding [9] requires very high computational cost (usually cost more than one second per frame and 16 seconds per clip), we encode raw depth into jet space (a fake color space) with three channels. We use jet encoding since it is more efficient to obtain comparable performance to using HHA encoding, actually, the computational cost of jet encoding is negligible compared to HHA encoding.

Without specific clarification, the results are reported as the overall accuracy (%) on video clips.

5.2 Evaluations on different conditions

In order to preliminarily evaluate the affects of lighting variance, we conduct the experiments with synthetic data that are generated to simulate different lighting conditions. The comparison results illustrated in Table 2 are obtained with the baseline 3D CNN models, since our goal is to evaluate the "pure" factors of conditions without affecting by training setting. Since depth data is not sensitive to the lighting conditions, only RGB synthetic data is generated for comparisons. While depth data is the source data captured with depth camera.

In Table 2, when only training CNN with ideally bright data, it achieves satisfactory results with ideal test data, yet obtains terrible results (closed to the results with random prediction) on dim and dark data. Such results suggest the significant affects of the lighting variance to RGB recognition. In addition, we also train 3D CNNs with dim and dark RGB data. Compared to the ideal data, training with dim or dark data obtains more robust results in different environments. It's relatively consistent that the best results are obtained when training and test data are in the same conditions, where training with ideal, dim and dark data obtains 55.1%, 47.1%, 43.9%, respectively. Particularly, we also make attempts to train CNN with the fusion of all data in ideal, dim and dark conditions. Although it can not obtain the better results in each single condition, it obtains the best mean result of all conditions. Note that the mean accuracy is also an important evaluation index to measure the robustness of scene recognition system in real world.

When combining with depth data (implemented with baseline RGB-D 3D architecture), the overall performances in different conditions are significantly improved. Particularly, training CNN with RGB-D ideal data outperforms only training with RGB ideal data in the largest margins, obtaining the gains of 4.3%, 28.2% and 26.0% on test of ideal, dim and dark data. These results particularly demonstrate the helpful power of depth data, particularly when training with ideal data, yet testing on data of aberrant conditions, such as dim or dark. The best mean result is obtained with the CNN trained the fusion of all data, and the variance between the results of different conditions is very slight, illustrating the robustness to different conditions. Compared to training with RGB data, training with the fusion of all RGB-D are more effective and robust, with the best mean accuracy of 54.9% and the smallest variance 0.6.

5.3 Evaluations with different attentions

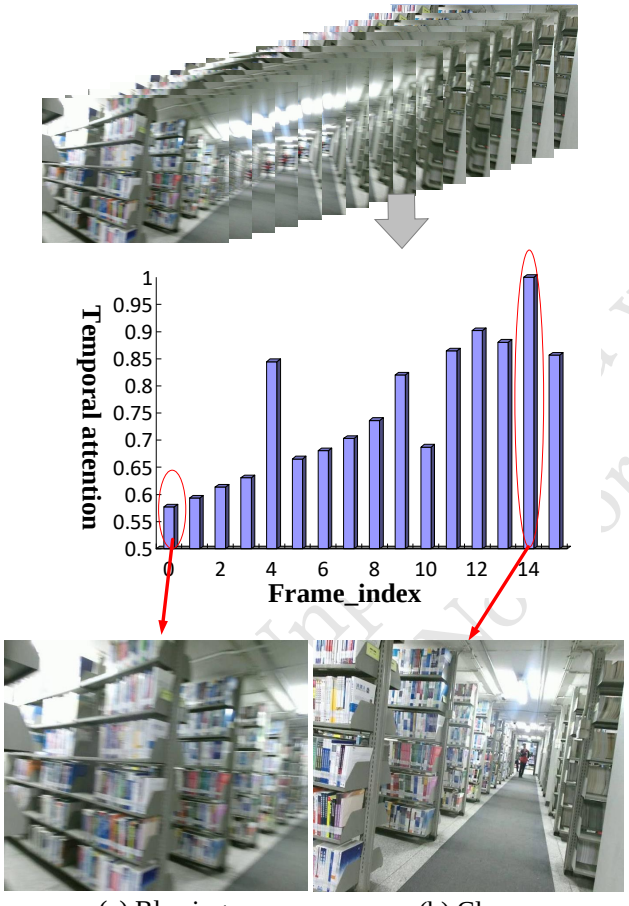
Consider different types of aberrant conditions, we particularly propose to integrate temporal, spatial and modal attentions into 3D CNN models. In order to evaluate the effects of proposed attentions, several corresponding experiments are conducted in following subsections.

5.3.1 Temporal attention for frame blurring. RGB frame blurring is unavoidable during the video capturing, yet it can be alleviated with more slowly moving. However, for the wide scenes, slower moving speed usually leads to larger capturing and processing time. Thus we propose to integrate temporal attention into CNN models to address blurring problem, allowing the higher speed capturing. In addition to ideal data, more challenging data is captured such as C-Fast and C-Faster, are integrated for evaluation, the comparison results are illustrated in Table 3.

When evaluating with ideal data, using the proposed temporal attention outperforms the baseline model with a gain of 2.3%. When evaluating with challenging data such as C-Fast and C-Faster, the improvements of the temporal attention are more significant, with the gains of 3.7% and 4.9%, respectively. With the baseline model, the results of C-Fast are better than C-Faster. While with the proposed temporal attention, we even obtain better results with C-Faster data, suggesting the effectiveness of temporal attention for

Table 2: Comparisons of different conditions in accuracy (%) with video clips, mean and variance of Idea, S-Dim and S-Dark are also reported

Bright	Training data			Modality		Test data				
	Ideal	Synthetic data		RGB	Depth	Ideal	Synthetic data		Mean	Variance
		S-Dim	S-Dark				S-Dim	S-Dark		
Depth specific	√	-	-	-	√	49.2	-	-	49.2	-
RGB specific	√	-	-	√	-	55.1	0.2	0.2	18.5	25.9
	-	√	-	√	-	30.0	47.1	41.7	39.6	7.1
	-	-	√	√	-	22.3	44.3	43.9	36.8	10.3
	√	√	√	√		40.9	42.8	43.2	42.3	1.0
RGB-D fusion	√	-	-	√	√	**59.4**	29.1	26.2	38.2	15.0
	-	√	-	√	√	46.1	55.4	54.6	52.0	4.2
	-	-	√	√	√	40.5	**56.3**	**55.2**	50.7	7.2
	√	√	√	√	√	55.5	55.2	54.1	**54.9**	**0.6**



(a) Blurring (b) Clear

Figure 4: Illustration of temporal attention on RGB data.

downplaying blurring frames. The key advantage of the proposed temporal attention is to improve the accuracy for faster capturing, allowing us to recognize wide scenes more efficiently.

One example (in test data of C-Faster) is also illustrated in Fig. 4, where a video clip with 16 frames captured in faster speed is shown in top, and the corresponding temporal (frame-wise) attention value is visualized in middle. Among the frames in this clip, the most blurring and clearest ones are emphasized in the bottom. It can be observed that rare details can be obtained with blurring frames, which should be typically downplayed.

Table 3: Evaluation of temporal attention in accuracy (%) with RGB video clips

Task	Methods	Test data	Accuracy
Ideal	Baseline [30]	Ideal	55.1
	Proposed	Ideal	**57.4**
Challenging	Baseline [30]	C-Fast	36.1
		C-Faster	35.4
	Proposed	C-Fast	**39.8**
		C-Faster	**40.3**

Table 4: Evaluation of spatial attention in accuracy (%) with depth video clips

Method	Modality	Data	Accuracy
Baseline [30]	Depth	Ideal	49.2
Proposed	Depth	Ideal	**50.6**

5.3.2 Spatial attention for depth vanishing. Spatial attention is particularly implemented for depth data, aiming to alleviate the affects of the depth vanishing regions. This problem is unavoidable to the video capturing during the camera moving. The comparison to the baseline model is illustrated in Table 4, where implementing the proposed spatial attention obtains a gain of 1.4% in accuracy. Weakening the depth vanishing regions with spatial attention is a compensative method to improve the error-tolerant rate during the camera moving in wide scenes. Spatial attention is necessary since the perfect capturing (camera) travel is not always ensured in real life scene video recognition.

5.3.3 Modal attention for lighting variance. Modal attention is implemented for the adaptive fusion of RGB-D models, aiming to dynamically emphasizing the modality that is resistant to aberrant conditions. Different types of data are involved in the comparisons,

Table 5: Evaluation of modal attention in accuracy (%) with RGB-D video clips, * indicates re-implementation by us

Method	Training data			Test data				
	Ideal	Synthetic data		Ideal	Synthetic data		Challenging data	
		S-Dim	S-Dark		S-Dim	S-Dark	C-Dim	C-Dark
Baseline [30]	√	-	-	59.4	29.1	26.2	36.7	20.9
	√	√	√	55.5	54.1	53.2	47.5	36.2
Proposed	√	-	-	**62.5**	28.6	26.9	40.5	20.0
	√	√	√	57.4	**56.1**	**54.4**	**49.5**	**39.5**
CNN-RNN* [28]	√	-	-	58.6	-	-	-	-

Table 6: Evaluations on full RGB-D videos in accuracy (%), * indicates re-implementation by us

Method	Training data			Test data				
	Ideal	Synthetic data		Ideal	Synthetic data		Challenging data	
		S-Dim	S-Dark		S-Dim	S-Dark	C-Dim	C-Dark
Baseline [30]	√	-	-	68.8	31.2	28.1	35.7	25.9
	√	√	√	64.8	63.4	62.4	50.3	44.3
Proposed	√	-	-	**72.8**	30.8	29.3	41.1	22.4
	√	√	√	66.7	**66.6**	**64.8**	**52.5**	**46.8**
CNN-RNN* [28]	√	-	-	68.5	-	-	-	-

which are illustrated in Table 5. When training with ideal data obtained in bright conditions, the proposed method outperforms the baseline with a gain of 3.1%/3.8 on test of ideal and C-Dim data, and obtain comparable results on other challenging tasks. Training with ideal data, the modal attention is not directly activated with lighting variance, but indirectly activated with blurring and depth vanishing. Since the dim and dark conditions may confuse the model trained with ideal data, it's reasonable that our model obtains comparable results to the baseline model.

When training with the merged data of ideal, S-Dim and S-Dark data, the proposed method outperforms the baseline on the tests of ideal, synthetic and challenging data. It's a particularly important insight that training with synthetic data is also effective to the recognition of "real" challenging data collected in real world. Training with synthetic data obtains the significant gains of 9.5% and 19.5% on the test of C-Dim and C-Dark compared to training with only ideal data. In contrast to addressing problems through capturing new data with brute force (to feed the CNN models), our goal is to alleviate the interference of challenging conditions with the proposed attention manners. Thus, without training with "real" challenging data, obtaining the robust scene recognition to different challenging conditions demonstrate the effectiveness of proposed method. Since RGB-D video scene recognition has been particularly researched in [28], which is the only related work for comparison. Our model outperforms [28] with a gain of 3.9%.

5.4 Evaluations on full videos

We also evaluate the proposed method with full videos, the comparison results are illustrated in Table 6. In general, the results with full videos are better than video clips, since one full video usually consists of several clips, more visual data captures more information. The proposed method also outperforms the baseline model on the tests of ideal, synthetic and challenging data. Particularly, training with merging of ideal and synthetic data significantly outperforms training with ideal data. Our model achieves the accuracy of 72.8% with ideal data, obtaining a significant gain of 4.3%, compared to the related RGB-D scene video recondition work [28]. For fair comparison, only the comparison of ideal data is reported, since [28] is not particularly designed for challenging conditions.

6 CONCLUSION

Scene recognition is to annotate scene labels for visual data such as images or videos. Previous works have made successes in scene recognition with the massive RGB data obtained from the ideal environments. However, scene recognition system in real world may dynamically face many types of aberrant conditions, damaging the performances of previous approaches. In contrast to previous works, our motivation is to train more robust CNN models to be adaptive to various aberrant conditions. In this paper, we propose an aberrance-aware framework for RGB-D scene recognition, which addresses the problems of RGB frame blurring, depth missing, and light variance by several types of attentions, such as temporal, spatial and modal attentions. All the attentions are homogeneously obtained with the gradient-sensitive maps of visual data, which are captured with the particularly designed convolutional kernels. In order to sense the gradients, different types of edge operators are implemented to design gradient based kernels. Finally, in order to evaluate the proposed methods, we make several attempts, such as obtaining more video from more challenging conditions and generating synthetic data to simulate different conditions, to conduct more sufficient comparisons. Particularly, we show that training the proposed CNN model with synthetic data can significantly improve the performances of scene recognition in more challenging conditions. Although capturing more "real" challenging data to feed CNN models may obtain more gains, we believe that addressing problems with the proposed methods is more efficient and robust.

REFERENCES

[1] Dan Banica and Cristian Sminchisescu. 2015. Second-Order Constrained Parametric Proposals and Sequential Search-Based Structured Prediction for Semantic Segmentation in RGB-D Images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] João Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. 4724–4733.

[3] Konstantinos G. Derpanis, Matthieu Lecce, Kostas Daniilidis, and Richard P. Wildes. 2012. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Jeff Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell, and Kate Saenko. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 2625–2634.

[5] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. 2007. What do we perceive in a glance of a real-world scene? *Journal of Vision* 7, 1 (2007), 10. DOI: https://doi.org/10.1167/7.1.10

[6] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. 2016. Spatiotemporal Residual Networks for Video Action Recognition. 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., 3468–3476. http://papers.nips.cc/paper/6433-spatiotemporal-residual-networks-for-video-action-recognition.pdf

[7] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. 2017. Temporal Residual Networks for Dynamic Scene Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[8] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[9] Saurabh Gupta, Pablo Arbelaez, Ross Girshick, and Jitendra Malik. 2014. Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation. *Int J Comput Vis* 112 (2014), 133–149.

[10] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2015. Indoor Scene Understanding with RGB-D Images: Bottom-up Segmentation, Object Detection and Semantic Segmentation. *International Journal of Computer Vision* 112, 2 (2015), 133–149. DOI: https://doi.org/10.1007/s11263-014-0777-6

[11] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. 2016. Cross Modal Distillation for Supervision Transfer. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[13] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1 (2013), 221–231.

[14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Fei-Fei Li. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. 1725–1732.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *NIPS*. 1106–1114.

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

[17] Fei Fei Li, Rufin VanRullen, Christof Koch, and Pietro Perona. 2002. Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences* 99, 14 (2002), 9596–9601. DOI: https://doi.org/10.1073/pnas.092277599

[18] Joe Yue-Hei Ng, Matthew J. Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond short snippets: Deep networks for video classification. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*. 4694–4702.

[19] A. Quattoni and A. Torralba. 2009. Recognizing indoor scenes. In *CVPR*.

[20] N. Shroff, P. Turaga, and R. Chellappa. 2010. Moving vistas: Exploiting motion for describing scenes. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 1911–1918. DOI: https://doi.org/10.1109/CVPR.2010.5539864

[21] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V (ECCV'12)*. Springer-Verlag, Berlin, Heidelberg, 746–760. DOI: https://doi.org/10.1007/978-3-642-33715-4_54

[22] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 568–576.

[23] K. Simonyan and A. Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations*.

[24] R. Socher, B. Huval, B. Bath, C. D. Manning, and A. Y. Ng. 2012. Convolutional-recursive deep learning for 3D object classification. In *NIPS*.

[25] Shuran Song, S. P. Lichtenberg, and Jianxiong Xiao. 2015. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*. 567–576. DOI: https://doi.org/10.1109/CVPR.2015.7298655

[26] Xinhang Song, Luis Herranz, and Shuqiang Jiang. 2017. Depth CNNs for RGB-D Scene Recognition: Learning from Scratch Better than Transferring from RGB-CNNs. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*. 4271–4277.

[27] Xinhang Song, Shuqiang Jiang, and Luis Herranz. 2017. Combining Models from Multiple Sources for RGB-D Scene Recognition. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*. 4523–4529.

[28] X. Song, S. Jiang, L. Herranz, and C. Chen. 2019. Learning Effective RGB-D Representations for Scene Recognition. *IEEE Transactions on Image Processing* 28, 2 (Feb 2019), 980–993. DOI: https://doi.org/10.1109/TIP.2018.2872629

[29] S. Thorpe, D. Fize, and C. Marlot. 1996. Speed of processing in the human visual system. *Nature* 381 (June 1996), 520–522. DOI: https://doi.org/10.1038/381520a0

[30] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features With 3D Convolutional Networks. In *The IEEE International Conference on Computer Vision (ICCV)*.

[31] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. 2016. Modality and Component Aware Feature Fusion For RGB-D Scene Classification. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[32] J. Xiao, J. Hayes, K. Ehringer, A. Olivia, and A. Torralba. 2010. SUN database: Largescale scene recognition from Abbey to Zoo. In *CVPR*.

[33] Jianxiong Xiao, A. Owens, and A. Torralba. 2013. SUN3D: A Database of Big Spaces Reconstructed Using SfM and Object Labels. In *Computer Vision (ICCV), 2013 IEEE International Conference on*. 1625–1632. DOI: https://doi.org/10.1109/ICCV.2013.458

[34] Yang Xiao, Jianxin Wu, and Junsong Yuan. 2014. mCENTRIST: A Multi-Channel Feature Generation Mechanism for Scene Categorization. *IEEE Trans. on Image Process.* 23, 2 (Feb 2014), 823–836. DOI: https://doi.org/10.1109/TIP.2013.2295756

[35] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. 2018. Places: An Image Database for Deep Scene Understanding. *IEEE Trans. on Pattern Anal. and Mach. Intell. (Accepted)* (2018).

[36] Hongyuan Zhu, Jean-Baptiste Weibel, and Shijian Lu. 2016. Discriminative Multi-Modal Feature Fusion for RGBD Indoor Scene Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.