DEPTH IMAGES COULD TELL US MORE: ENHANCING DEPTH DISCRIMINABILITY FOR RGB-D SCENE RECOGNITION

Dapeng Du, Xiangyang Xu, Tongwei Ren, Gangshan Wu

State Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
dudp.nju@gmail.com,xiangyang.xu@smail.nju.edu.cn, rentw, gswu@nju.edu.cn

ABSTRACT

Recently depth-modal information has been witnessed effectively in computer vision community, especially for scene analysis related tasks. However, it still suffers severely from depth data scarcity as well as improperly transferring pre-trained RGB models to fit depth-modal data. In this study, we propose a novel two-step training strategy to address these problems and focus on enhancing the recognition power for depth-modal images in RGB-D scene recognition task. Specifically, we build an effective "Res-U" architecture on a GAN (generative adversarial networks) based RGB-to-depth modality translation model, which is endowed with both short and long skips for residual learning. On one hand, this could first well pre-train a depth-modal-specific discriminator network from scratch in an unsupervised manner, which is effectively transformed for the subsequent recognition task instead of directly fitting pre-trained RGB model to depth-specific one. On the other hand, new depth images with helpful perturbations, generated from the modality translation model, help argument the original training set and regularize the learning process in some sense. This two-step training strategy makes it more effective for training a modal-specific network to discriminate depth scenes. Besides, we extensively explore the modality translation network to investigate the effects in recognizing depth-modal scenes, which encourages a reasonable way to take full advantage of multi-modalities. The proposed method achieves state-of-the-art accuracy on NYU Depth v2 and SUN RGB-D benchmark datasets, especially on depth data only evaluation.

Index Terms— RGB-D image, scene recognition, indoor scene

1. INTRODUCTION

Recent years have seen kinds of visual tasks leveraging depth images captured by RGB-D cameras, which have especially achieved significant performance with CNN architectures, *e.g.*, object detection [2], image segmentation [3], saliency detection [4], etc. As a complementary modality for RGB images, depth images are less insensitive to illumination changes and texture variation. Notably in scene analysis related tasks, depth images are capable of providing clearer and more intuitive spatial description of the captured scenes, so that the tasks have been propelled by a big step in the last few years [5]. However, there still exist some obvious, yet important problems. On one hand, CNN based methods were suffering from labeled depth data scarcity to show its power. In spite of the significant scale growth of dataset achieved by the recent SUN RGB-D dataset [5], which helps step up the pace for RGB-D related research, its order of magnitude is still much smaller than existing large-scale RGB datasets (e.g., ImageNet [1], Place dataset [6]). Training depth-modal model from scratch using depth data at such scale would often cause severe data over-fitting problem. Besides, though transfer learning or model reuse, which attempts to construct a model by utilizing existing pre-trained model rather than build it from scratch, provides certain help for this target data limited situation, we found that the features learned from fine-tuning pre-trained RGB models with depth-modal data are obviously not depth-specific, as shown in Fig. 1. We observed that even if a pair of depth and RGB images are two perspec-

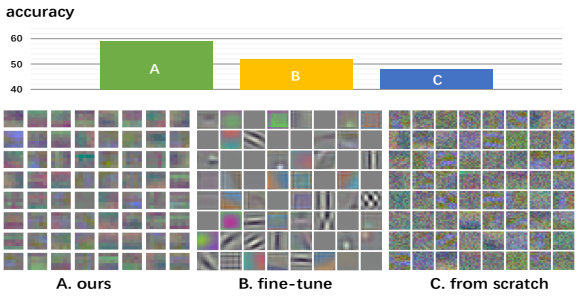

Fig. 1. Examples of how CNN models could affect the depth-modal recognition performance by visualizing the first convolutional layer (Conv1). All three models use similar architectures training with depth images where A is from our proposed depth-modal-specific recognition network while B is pre-trained on ImageNet [1] in advance then fine-tuned with depth images. C trains from scratch. Limited depth data would make it hard to truly train satisfactory model either by fine-tuning from RGB models or directly training from scratch.

978-1-5386-1737-3/18/$31.00

tives of the same scene exhibiting the similar contour and semantics, there exist intrinsic differences between these two modalities, thus it would fail to truly learn good depth-modal-specific features by directly fitting RGB-specific model to depth-specific one. The forementioned above inspires us to wonder if we could get around the problem by directly training a depth-modal model using limited depth data but would still be effective for this task.

In this study, we will focus on improving the discriminative power of depth model to address the RGB-D scene recognition problem. The basic idea is that even the training data is limited, if we properly train a neural network for one task at first, then extend and fine-tune it with another task (*i.e*., by different loss functions) on the same data, the network would still be able to learn more discriminative features for latter tasks. Specifically, we train the depth recognition network using a two-step strategy. We first introduce an effective "Res-U" architecture endowed with both short and long skips for residual learning building on a GAN based RGB-to-depth modality translation model. This could first well pre-train a depth-modal-specific discriminator network from scratch in an unsupervised manner, which is further effectively transformed to a scene recognition network using cross-entropy loss. This could avoid improperly fine-tuning from RGB models but still get a well-trained depth-modal network by training with two different losses. What's more, photo-realistic auxiliary depth scenes with help perturbations are generated to augment original training set, which in some sense alleviates the depth-modal data scarcity. We also extensively explore the modality translation network to investigate the effects on depth scenes recognition, which encourages a reasonable way to take full advantage of multi-modalities.

To evaluate the performance of the proposed method, we conduct extensive experiments on two challenging RGB-D benchmark datasets, NYU Depth v2 [7] and SUN RGB-D [5]. The experimental results show that the proposed method outperformed the start-of-the-art methods both on RGB-D and depth modality only evaluations. The main contributions of our study are as follows.

- We propose to approach the RGB-D indoor scene recognition problem on training a more discriminative depth-modal model. We propose a novel two-step strategy to train a depth-modal-specific network from scratch instead of directly fitting pre-trained RGB model to depth-specific model, which effectively improves the discriminability on depth scenes, thus improve the final performance by a significant margin.
- In the modality translation model, we introduce an effective "Res-U" architecture endowed with both short and long skips for residual learning to well-train a depth-modal-specific network. More over, we extensively explore the modality translation network to investigate the effects for training the subsequent depth-modal-specific recognition network, which encourages a reasonable way to take full advantage of multi-modalities.

2. RELATED WORK

RGB-D Scene Recognition. Currently many studies explore CNN features for RGB-D indoor scene recognition to avoid the limitations of hand-craft features [8, 9]. In [10], the correlation of inter- and intra-modality on RGB-D CNN features was explicitly modeled for all scene categories. Wang *et al*. [11] fused the proposal based Fisher Vector (FV) [6] features embedded jointly from RGB, HHA (Horizontal disparity, Height above ground and Angle with gravity) [2] and surface normal to perform scene classification. These methods rely on transferring pre-trained CNN on Places dataset [12] to fine-tune RGB and depth image pairs. Different from these studies, we focus on improving the discriminative power of depth model directly without relying on fine-tuning from pre-trained RGB models. In [13], Song *et al*.argued that the depth features fine-tuned from large-scale labeled RGB datasets are biased and tried to learn depth features from scratch using depth patches. However, they totally sacrificed the benefits from RGB models on depth-modal network training while our method is obviously different from that since we use a modality translation model to pre-trained the depth-modal recognition network in an unsupervised manner which takes full advantage of multi-modalities. The proposed depth-modal recognition network is relatively simpler but could discriminate depth scenes more effectively.

Scene Generation with GAN. GANs [14, 15, 16, 17, 18] have achieved impressive results in image generation. Makhzani *et al*. [15] incorporated generative adversarial networks (GAN) [14] in auto-encoders to perform variational inference. Radford *et al*. [16] proposed a deep convolutional generative adversarial networks (DCGANs) to generate interpretable images with richer information, such as faces and bedrooms. Xiaolong Wang *et al*. [17] factorized the image generation process and proposed Style and Structure Generative Adversarial Network(S^2-GAN). These works mostly either focused on generating RGB images or used randomly sampled noises to generate images, which limit the reuse of learned representations. Recently, in [18], the authors proposed to use GANs to learn the mapping functions between two image domains with images as input instead of random noise vectors, known as image-to-image translation. In this study, we build our work upon [18]. Unlike most previous work only focusing on the effects of generated images, we also take great interest in the discriminator network which is specially extended to the depth-modal-specific recognition network to effectively discriminate depth scenes.

3. METHOD

In our learning procedure for RGB-D indoor scene recognition, we first present the two-step strategy for training a

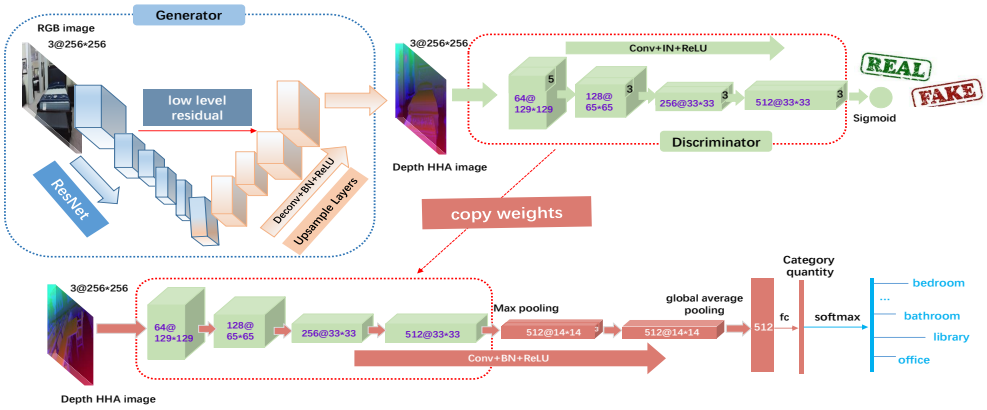

Fig. 2. The proposed procedure for learning depth-modal recognition network. Cubes are feature maps, with dimensions represented as #features@height∗width. Note that for the upper Discriminator network, we use Instance Normalization (IN) for image generation task and when transforming it to the subsequent depth recognition network, a Batch Normalization (BN) is added after the last Convolution operation for classification task.

depth-modal-specific model which could take full advantage of multi-modalities without suffering from biased features from pre-trained RGB models, then inspect the proposed architecture in this task. Consequently, we discuss the synthetic depth images used for training. Similar to [2, 5], we employ geocentric encoding of depth images, HHA (Horizontal disparity, Height above ground and Angle with gravity) embedding[1], to capture the scenes structural and geometrical properties.

3.1. Reprocessing from GAN Loss

There are two parts with specific networks in the procedure of training the depth-modal recognition model. The first part is a conditional GAN model adapted from [18] where the generator network is an encoder-decoder structure to map high resolution RGB input images to according output depth images; the discriminator gives a yes or no to real and generated depth scenes in an unsupervised way, *i.e.*, using an adversarial loss [14]:

$$L_{adv}(\mathcal{G}_D, \mathcal{D}_D, x, y) = \min_{w_{\mathcal{G}_D}} \max_{w_{\mathcal{D}_D}} \sum_{y} \log \mathcal{D}_D(y; w_{\mathcal{D}_D}) + \sum_{x} \log(1 - \mathcal{D}_D(\mathcal{G}_D(x; w_{\mathcal{G}_D}); w_{\mathcal{D}_D})) \quad (1)$$

where $w_{\mathcal{G}_D}$ and $w_{\mathcal{D}_D}$ are weights to be learned for the generator and discriminator, respectively. This unsupervised learning procedure tries to learn true data distribution for generator to produce more genuine samples as well as the anti-fooled ability for the discriminator.

The GAN part network is mainly intended to train a relatively "good" D network[2] using an unsupervised adversarial loss (see Eq. 1) and produce extra synthetic depth images to augment the original training set in passing. We regard this as a way of taking advantages of multi-modalities for pre-training a depth-modal network. When GAN loss has been plateau, we transform this pre-trained D network to a depth recognition network for scene classification task. Our insight is that by applying another different task (loss) to guide a pre-trained network, the Conv units would still have room to learn better presentation from the same training data for latter tasks. More specifically, we copy the CNN layers of the D network and transform it to the recognition network by adding max-pooling, normalization and fully-connected (fc) layers. Since the complexity of depth images is much lower than that of RGB images (e.g., no textures or illumination changes), we use global average pooling for the last Conv feature maps and impose only one fc layer to reduce the total parameters of the network. The whole architecture is illustrated in Fig. 2.

3.2. Investigating Generator Network

Unlike many studies which mainly focus on the performance of G network, we try exploring how to take better advantage of RGB models for training the depth-modal discriminator and further improve the performance of depth-modal recognition network. Since the bottom layers of G network are in some sense in charge of extracting RGB features while the top layers care more about depth-modal cues, it would go on more smoothly for training G network if a better RGB model (*e.g.*, pre-trained on a more relative large-scale dataset) is available. Accordingly we could expect the discriminator get better trained due to the "two-player game" mechanism and further improve the latter recognition task.

[1]We use the term depth and HHA interchangeably, if not specified.

[2]G network and D network refer to Generator network and Discriminator network in generative adversarial networks respectively.

Table 1. Ablation study for different models on depth network, tested on NYU Depth v2 test set (accuracy %).

Models	Original training set	Gen depth added
SC	54.4	-
FT	55.1	-
Res-SC	55.2	57.0
Res-FT	56.8	58.3
Res-U	57.8	59.0

Table 2. Comparison on different losses for depth-modal scene recognition on the NYU Depth v2 test set and SUN RGB-D test set in terms of classification accuracy (%). Training procedure is without including generated depth images.

Loss	NYU Depth v2	SUN RGB-D
ℓ_1	55.1	40.7
Huber	56.6	42.2

Toward this goal, we build the downsample layers of the G network on ResNet [19], which introduces skip layers that by-pass two or several convolutions being summed up to their output. Considering that there is still a lot of common low-level information shared between RGB and depth images (e.g., contours and edges), despite residuals by short steps in ResNet network, it would be desirable to pass long-term information across the net directly. To this end, we add one skip connection, from the first Conv layer of the ResNet to the last Conv layer of the upsample network to make it a U-shape [20]. Thus this proposed "Res-U" network is endowed with two types of skip layers for different aspects of considerations. Instead of qualitatively checking the quality of generated scenes, we quantitatively evaluate the subsequent scene recognition performance with this setting compared to several ResNet baselines. The result is illustrated in Table. 1.

For image-to-image translation GAN model, there usually exists a constraint loss for correspondence of input and output pair on visual appearance and spatial structure, *e.g.*, the ℓ_1 loss used in [18]. We also explore using a different constraint on the consideration of not only the performance of the generated depth images but is also beneficial to subsequent recognition task. We attempt the Huber criterion [21] (see, Eq. 2, here x refers to $\mathcal{G}_D(x; w_{\mathcal{G}_D})$ for simplicity) for a simple but effective intuition that the ground truth of depth map usually has many outliers because of the equipment's characteristic and we found Huber criterion is less sensitive to these outliers and in some cases prevents exploding gradients (*e.g.*, more details could be found in [22]). Table 2 quantifies the performance on using Huber loss function and ℓ_1 loss function respectively. It turned out that the Huber criterion is consistently superior over the ℓ_1 by a observable margin.

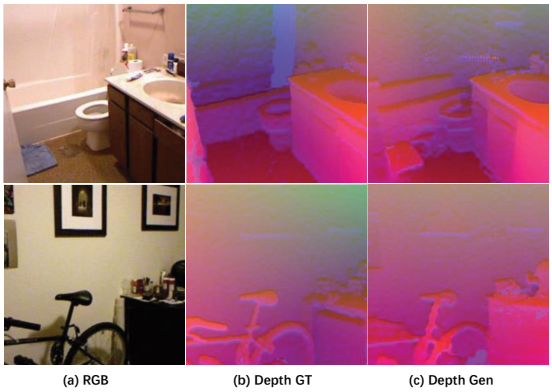

Fig. 3. Exemplars of generated depth images on NYU Depth v2 test set. Column (a) consist of the ground-truth RGB images, (b) are the corresponding ground-truth depth maps, and (c) are the generated ones.

All forementioned experiments indicate that by better adjusting the generator we could reasonably obtain better results in depth-modal recognition task. More details could be found in ablation study in Sec.4.3.

$$L_{rec}(x,y) = \begin{cases} \frac{(x-y)^2}{2}, & |x-y| < 1 \\ |x-y| - \frac{1}{2}, & \text{otherwise} \end{cases} \quad (2)$$

3.3. Training with Generated Depth Images

The proposed method comes with another advantage. The newly generated depth images that are located near the real training ones introduce more perturbations in terms of deformation or spatial layout, which can regularize the learning process in some degree. Fig. 3 shows an example of generated depth images. This also encourages the depth-modal recognition network to be less prone to over-fitting. Therefor, we augment the original training set of the depth images with the newly generated depth ones, which are assumed as the same scene category as the input RGB image.

4. EXPERIMENTS

4.1. Dataset

We comprehensively evaluate our approach on two popular RGB-D benchmark datasets, the NYU Depth v2 [7] and SUN RGB-D [5]. The former is comprised of 1449 densely labeled pairs of aligned RGB and depth images while the SUN RGB-D contains totally 10335 ones. We use 795 / 654 for training / testing splits on NYU Depth v2 and 4845 / 4659 on SUN RGB-D; the splits are grouped into nine most common categories with an "others" category for the NYU Depth v2 set and nineteen major categories for the SUN RGB-D set respectively, as per the standard splits and experiment settings stated in [5, 11]. To measure the RGB-D scene recognition, we quantitatively report the mean precision (classification accuracy) over all scene categories in both datasets.

Table 3. Comparison with the state-of-the-art methods on the NYU Depth v2 test set in terms of classification accuracy (%).

Method	RGB	Depth	RGB-D
Baseline	53.4	51.8	59.5
Gupta *et al*.[23]	-	-	45.4
Wang *et al*.[11]	53.5	51.5	63.9
Zheng *et al*.[24]	-	-	61.4
Song *et al*.[13]	53.4	56.4	65.8
Song *et al*.[25]	-	-	66.7
Ours	53.7	**59.0**	**67.5**

Table 4. Comparison with the state-of-the-art methods on the SUN RGB-D test set in terms of classification accuracy (%).

Method	RGB	Depth	RGB-D
Baseline	41.5	37.5	45.4
Liao *et al*.[26]	36.1	-	41.3
Zhu *et al*.[10]	40.4	36.5	41.5
Wang *et al*.[11]	40.4	36.5	48.1
Song *et al*.[13]	42.7	42.4	52.4
Song *et al*.[25]	-	-	52.3
Ours	42.6	**43.3**	**53.3**

4.2. Architecture Details and Experiment Setup

All the ResNet architectures employed in this paper are similar to ResNet-18 [19]. For the GAN part learning rate is set to 2×10^{-4} for the first 100 epochs and linearly decays over the next 100 epochs. In the procedure of training the recognition network, we fine-tune the reorganized D network with a smaller learning set as 2.5×10^{-5} also with a linear decay strategy. The RGB and depth networks are fused by connecting from the last feature layer, concatenating their respective features to a new fc layer for integration. This makes it a end-to-end learning network which could simultaneously fine-tune two modality networks for better recognition performance.

4.3. Evaluations on the NYU Depth v2 Set

We first make an ablation study of our method with five conditions separately evaluated on NYU Depth v2 depth data, *i.e*., the depth recognition network is trained: 1) From Scratch in a direct way (SC); 2) Pre-trained on ImageNet then fine-tuned with depth images (FT); 3) Leveraging GAN network to pre-train a discriminator with the downsample part (ResNet) of G network training from scratch (Res-SC); 4) Similar with 3) but the downsample network is pre-trained on ImageNet (Res-FT); 5) Leveraging the "Res-U" architecture to pre-train a discriminator (Res-U). Note that we only evaluate the depth-modal recognition performance here. Experiments 3) - 5) demonstrate the necessity of training by the proposed two-step strategy compared with 1) - 2), their inner different settings exhibiting the effects of adjusting the generator network. The result is illustrated in Table 1. We can see that directly fine-tuning a RGB model with depth images is better than training from scratch, but not good enough. Training with the proposed two-step strategy obviously shows its effectiveness and the experiment results encourage a reasonable way, *e.g*., choosing a better pre-trained RGB model in downsample part, to take full advantage of multi-modalities for enhancing the recognition power of depth-modal-specific network. Moreover, by integrating the newly generated depth scenes, we observe a further improvement about 2%. We give the credit to the helpful perturbations brought from the newly generated depth images, which in some sense could be treated as a more natural regularization during the training procedure, thus alleviate the over-fitting problem.

Table 3 shows comparisons with state-of-the-art methods. We compare with: 1) Gupta *et al*. [23] leveraged segmentation responses as features for scene recognition; 2) Wang *et al*. [11] fused RGB, depth, and surface normals Fisher Vector features, which are encoded from proposal-based CNN features; 3) Zheng *et al*. [24] exploited multi-task metric multi-kernel learning on off-the-shelf multimodal CNN features; 4) Song *et al*. [13] developed weakly supervised patch CNN to learn depth-specific features; 5) Song *et al*. [25] combined multi-source depth features for scene recognition, which was built upon one RGB network and two depth networks. Baseline method is also from [13]. The proposed method outperformed others on final RGB-D performance, notably on depth modality only evaluation by a big margin, which indicates that our method learned better depth-modal-specific features for recognizing depth-modal scenes.

4.4. Evaluations on the SUN RGB-D Set

We compare our proposed method to the state-of-the-art on SUN RGB-D benchmark in Table 4. We compare with: 1) Liao *et al*. [26] incorporated features extracted from semantic segmentation to improve scene recognition; 2) Zhu *et al*. [10] took inter- and intra-modalities correlation into consideration for RGB-D scene recognition; 3) Wang *et al*. [11]; 4) Song *et al*. [13]; 5) Song *et al*. [25]. Similarly, baseline method is from [13]. This dataset is much more complicated since images are captured by various RGB-D sensors and there exists sensor bias due to the diverse capabilities for different sensors [5]. Therefore, it is important for methods to generalize to adapt to different types of RGB-D sensors' capturing characteristics. Under this condition, our proposed method still performed robustly and achieved the best performance on depth only and the final RGB-D evaluations.

5. CONCLUSION

In this paper, we presented an effective method with two-step training strategy on enhancing depth-modal recognition power for RGB-D indoor scene recognition task. By leveraging a GAN based RGB-to-depth translation model with an effective "Res-U" architecture, helpful photo-realistic depth images are generated and a reasonable depth-modal-specific network is pre-trained in a unsupervised way. Moreover, we also explore how to take full advantage of multi-modalities for modal-specific network, which encourages a reasonable way to enhance the recognition power for the depth-modal network. In the future, we will further explore the image-to-image translation model on reversely using the relatively less complicated depth data to help improve the RGB recognition network.

6. REFERENCES

[1] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *IJCV*, vol. 115, no. 3, pp. 211–252, 2015.

[2] Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik, "Learning rich features from rgb-d images for object detection and segmentation," in *ECCV*. Springer, 2014, pp. 345–360.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014, pp. 580–587.

[4] Congyan Lang, Tam V Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan, "Depth matters: Influence of depth cues on visual saliency," in *ECCV*, pp. 101–115. Springer, 2012.

[5] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015, pp. 567–576.

[6] Jorge Sánchez, Florent Perronnin, Thomas Mensink, and Jakob Verbeek, "Image classification with the fisher vector: Theory and practice," *IJCV*, vol. 105, no. 3, pp. 222–245, 2013.

[7] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus, "Indoor segmentation and support inference from rgbd images," *ECCV*, pp. 746–760, 2012.

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

[9] Jianxin Wu and Jim M Rehg, "Centrist: A visual descriptor for scene categorization," *PAMI*, vol. 33, no. 8, pp. 1489–1501, 2011.

[10] Hongyuan Zhu, Jean-Baptiste Weibel, and Shijian Lu, "Discriminative multi-modal feature fusion for rgbd indoor scene recognition," in *CVPR*, 2016, pp. 2969–2976.

[11] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham, "Modality and component aware feature fusion for rgb-d scene classification," in *CVPR*, 2016, pp. 5995–6004.

[12] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva, "Learning deep features for scene recognition using places database," in *Advances in neural information processing systems*, 2014, pp. 487–495.

[13] Xinhang Song, Luis Herranz, and Shuqiang Jiang, "Depth cnns for rgb-d scene recognition: Learning from scratch better than transferring from rgb-cnns.," in *AAAI*, 2017, pp. 4271–4277.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[15] Alireza Makhzani, Jonathon Shlens, Navdeep Jaitly, Ian Goodfellow, and Brendan Frey, "Adversarial autoencoders," *arXiv preprint arXiv:1511.05644*, 2015.

[16] Alec Radford, Luke Metz, and Soumith Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[17] Xiaolong Wang and Abhinav Gupta, "Generative image modeling using style and structure adversarial networks," in *European Conference on Computer Vision*. Springer, 2016.

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," *arXiv preprint arXiv:1611.07004*, 2016.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.

[20] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2015, pp. 234–241.

[21] Peter J Huber, "Robust statistics," in *International Encyclopedia of Statistical Science*, pp. 1248–1251. Springer, 2011.

[22] Ross Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[23] Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik, "Indoor scene understanding with rgb-d images: Bottom-up segmentation, object detection and semantic segmentation," *IJCV*, vol. 112, no. 2, pp. 133–149, 2015.

[24] Yu Zheng and Xinbo Gao, "Indoor scene recognition via multitask metric multi-kernel learning from rgb-d images," *Multimedia Tools and Applications*, vol. 76, no. 3, pp. 4427–4443, 2017.

[25] Xinhang Song, Shuqiang Jiang, and Luis Herranz, "Combining models from multiple sources for RGB-D scene recognition," in *IJCAI 2017, Melbourne, Australia, August*, 2017, pp. 4523–4529.

[26] Yiyi Liao, Sarath Kodagoda, Yue Wang, Lei Shi, and Yong Liu, "Understand scene categories by objects: A semantic regularized scene classifier using convolutional neural networks," in *ICRA*. IEEE, 2016, pp. 2318–2325.