Translate-to-Recognize Networks for RGB-D Scene Recognition

Dapeng Du   Limin Wang*   Huiling Wang   Kai Zhao   Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, China

Abstract

Cross-modal transfer is helpful to enhance modality-specific discriminative power for scene recognition. To this end, this paper presents a unified framework to integrate the tasks of cross-modal translation and modality-specific recognition, termed as Translate-to-Recognize Network (TRecgNet). Specifically, both translation and recognition tasks share the same encoder network, which allows to explicitly regularize the training of recognition task with the help of translation, and thus improve its final generalization ability. For translation task, we place a decoder module on top of the encoder network and it is optimized with a new layer-wise semantic loss, while for recognition task, we use a linear classifier based on the feature embedding from encoder and its training is guided by the standard cross-entropy loss. In addition, our TRecgNet allows to exploit large numbers of unlabeled RGB-D data to train the translation task and thus improve the representation power of encoder network. Empirically, we verify that this new semi-supervised setting is able to further enhance the performance of recognition network. We perform experiments on two RGB-D scene recognition benchmarks: NYU Depth v2 and SUN RGB-D, demonstrating that TRecgNet achieves superior performance to the existing state-of-the-art methods, especially for recognition solely based on a single modality.

1. Introduction

Recently computer vision community has concentrated on applying convolutional neural networks (CNN) [15] to various vision tasks [10, 22, 8, 28, 30, 29]. Meanwhile, the rapid development of cost-affordable depth sensors (*e.g.*, Microsoft Kinect and Intel Realsense) have triggered more attractions to revisit computer vision problems using RGB-D data, such as object detection [7, 34], image segmentation [19], activity recognition [36, 31].

In this paper, we focus on enhancing the modality-specific network's representative power of RGB-D scene recognition, the goal of which is to accurately classify the

---

*Corresponding author.

given scene images with aligned color and depth information. It is a highly challenging classification task on understanding RGB-D scene data on one hand for its essentially diverse combination of cluttered objects, bounding in different semantic representations for indoors scenes, and importantly on the other hand for the data scarcity problem. The amount of the largest RGB-D dataset [25] is still order-of-magnitude insufficient to provide enough labeled RGB-D data. Since pre-trained RGB CNN models are easily adapted for new RGB data, recent works have focused on learning effective depth features. A few methods [27, 40] directly use the pre-trained RGB CNN weights to fine tune a depth CNN, but only limited improvements are given. [7] directly transferred semantic supervision from labeled RGB images to unlabeled depth images, which has limits on transform direction.

This paper tackles major challenges mentioned above from two aspects: 1) We propose to enhance the discriminative power for both RGB and depth's single-modal networks by a cross-modal translation procedure, and 2) we enhance training data sampling with the generated images of high semantic relevance for the classification task. The basic idea is that the modality translation enhances the description power of the encoding network as it forces the RGB/depth data to infer information towards its complementary modality. RGB$\rightarrow$ depth translation network could improve the representation ability of RGB scene network, by learning generating depth data which encodes better on geometric and appearance invariant cues, while depth scene network get benefited learning color and texture cues through depth$\rightarrow$ RGB translation. Meanwhile, this translation process produces new cross-modal data of high quality for the other modality data's classification task.

Specifically, we propose to couple an arbitrary modality-specific scene **Recog**nition network with a modality **Tr**anslation network trained in a multi-task manner, termed as **TRecg**Net; both branches share the same Encoding Network (*E* Net), as shown in Figure 1. The effectiveness of TRecgNet that boost modality-specific description power essentially lies on the effect of modality translation, *i.e.*, how the RGB-D data could effectively learn the semantic similarity from the paired data to boost the jointly learn-
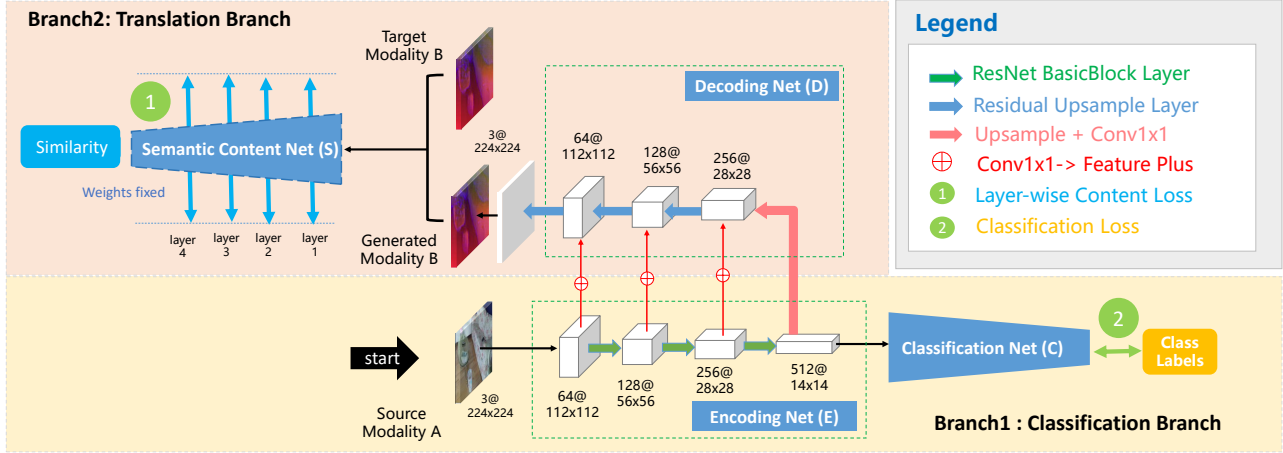
Figure 1. **TRecgNet.** Cubes are feature maps with dimensions and size represented as #features@height∗width. The pipeline consists of two parallel streams: 1) **Recognition Branch** is for recognizing scene images, in which $E \rightarrow C$ Nets are updated using supervised classification loss. 2) **Translation Branch** aims at constructing paired complementary-modal data of input through $E \rightarrow D$ Nets. The translation procedure is constrained by semantic supervision S Net. We jointly train the two branches in an end-to-end manner. In the test phase, we only use the Recognition Branch.

ing task. We do not simply use the pixel-level Euclidean distance loss as the supervision like many reconstruction works did [35]. We argue that this difference is unreliable, especially for RGB → depth translation, since the ground truth images of depth data usually exist many outliers because of the equipment's limitations and operation errors. Also, low level pixel-wise similarity fails to provide any semantic relevance. Instead, we perform the translation process using pre-trained semantic model, which is inspired by the style transfer related works [3, 13, 2]. It has been shown that CNNs trained with sufficient labeled data on specific vision tasks, such as object classification, has already learned to extract semantic content representations. This generalized ability is not limited to specific datasets or tasks. In these works, authors use the perceptual constraint from one specific layer, usually conv4_x of VGG model, to keep the "rough" content of the image when transforming it to another style. In contrast, we propose to leverage perceptual loss from multiple layers supervising the translation process, for a simple yet effective intuition that higher layer tends to preserve the semantic content while lower one does better in capturing the detailed information, which could provide enough and effective cues for cross-modal similarity learning.

We test the TRecgNet on the two benchmarks of the RGB-D Indoor scene recognition task, SUN RGB-D dataset and NYU Depth dataset v2. Our TRecgNet could obtain an evident improvement on both modality-specific and RGB-D settings. The main contributions of this paper are two-fold:

- A TRecgNet is proposed to transfer complementary

cues through a label-free modality translation process, which improves modality-specific classification task in an end-to-end manner and achieves state-of-the-art performance on RGB-D indoor scene benchmarks.

- TRecgNet can generate more photo-realistic and semantically related data to enhance training data sampling which alleviates the data scarcity problem and effectively improves the classification performance.

2. Related Work

RGB-D Scene Recognition. Earlier works relied on handcrafted features to capture the characteristic properties of the scene. Banica *et al*. [1] used second-order pooling to quantize local features for segmentation and scene classification. Gupta *et al*. [5] proposed to quantize segmentation outputs by detecting contours from depth images as local features for scene classification. Recently, multi-layered networks such as CNNs is able to learn useful representations from large amounts of data. In general, they learned modality-specific features from RGB and depth images separately, and then performed the fusion. Wang *et al*. [27] extracted and combine deep discriminative features from different modalities in a component aware fusion manner. Gupta *et al*. [7] transferred the RGB model to depth net using unlabeled paired data according to their mid-level representations. However, these methods only consider transferring color cues to depth net and ignore that depth cues could also benefit the RGB net. More recently, the recognition power of depth net has been comprehensively studied. In [26], Song *et al*. argued that learning depth features from pre-trained RGB CNN models are biased and seek to learn

depth features from scratch using weakly supervised depth patches. However, since depth data still suffers from data scarcity problem, the structure of the depth nets fail to go deep which limits its extendibility .

Paired Image-to-Image Translation. Paired image-to-image translation problems could be formulated as pixel level mapping functions, however, each pixel is translated independently from others in this manner [37]. Recently, GANs [4, 21] have achieved impressive results in image generation. Wang *et al.* [32] factored the generation into two steps: surface normal generation and content generation. They used a pixel-wise surface normal constraints as additional supervision. In [18], the authors proposed to use GANs to learn the mapping functions between paired images. These GAN constraints help generate more variegated images from data distribution learning, however, GAN based model is hard to train and generated images tend to lack the semantic consistency of source images that are hard to be leveraged in subsequent semantics related work, such as classification or segmentation tasks. More recently, in many style transfer works [3, 13, 2], they use perceptual loss from specific layer of pre-trained RGB models, to maintain the structural content during the translation. However, most of them only use content constraint from one specific layer, usually conv4_x of VGG model, ignoring that pre-trained models could perform different levels of semantic supervisions which are good enough for image translation. In this paper, we propose to leverage perceptual loss from multiple layers to constrain the translation process of RGB-D data. Unlike pure image generation task, we focus on how this translation process finally benefits the classification task and the performance of data augmentation using generated images.

3. Method

This section details the description of our Translate-to-Recognize Network (TRecgNet). The proposed framework is illustrated in Figure 1. Assume that in RGB-D setting we would like to train a classification network using one modality data. Let $(\mathbf{M_A}, \mathbf{M_B})$ be the paired RGB-D images of specific classes from set $L = \{1, 2, ..., N_c\}$, where N_c is the total number of scene classes. Our object is to learn an embedding $E : \mathbf{M_A} \rightarrow \mathbb{R}^d$ with a translation mapping $T : \mathbb{R}^d \rightarrow \mathbf{M_B}$ and a class prediction function $C : \mathbb{R}^d \rightarrow \mathbf{L}$. The core problem is how the translation process could make the modality-specific Encoder Network (E Net) learn effective complementary-modal cues to benefit the classification task.

3.1. The TRecgNet Architecture

TRecgNet consists of four parts including Encoder Network (E Net), Classification Net (C Net), Decoder Net (D Net) and Semantic Content Network (S Net). Figure 1
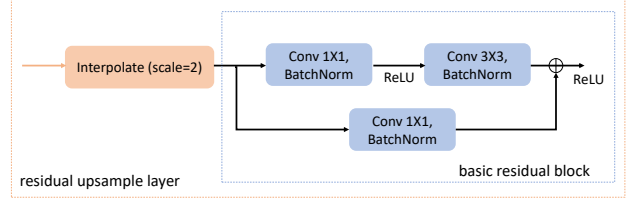

Figure 2. Architecture of the residual upsample layer. The input feature maps get upsampled with scale 2 followed by one basic residual block.

shows the exemplar architecture of RGB TRecgNet built on ResNet-18, which uses pre-trained ResNet18 as the E and C Nets. D decodes the feature maps from E to reconstruct the complementary data. Depth TRecgNet uses the same structure only by exchanging the position of source and target modalities. We make three measures to enhance the translation process. First, we empirically remove the first max-pooling operation of ResNet. In the whole structure, feature maps only shrink by convolution operation with stride and less information would get lost which is very important for image translation process. Second, we introduce residual upsample layer in the D Net. A residual upsample layer upsamples feature maps with a bilinear interpolation operation accompanied with one residual block which mimics the basic residual block of ResNet. Figure 2 shows the architecture of the residual upsample layer. Third, we propagate the context information from three stages of E to the corresponding outputs of D similar with [23]. We use the plus operation instead of concatenation which could reduce the number of parameters in the D Net.

3.2. Layer-wise Content Supervised Translation

The translation network aims to improve the presentation ability of E Net learning characteristics of complementary data, the procedure of which is supervised by a semantic RGB CNN model, which uses the combination of perceptual constrains from lower layers to higher ones, measuring the **layer-wise similarity** of the generated and the paired data.

We use the ResNet model pre-trained on ImageNet [38] as the supervision content network S, for the consideration of accordance with the architecture of E Net. More details could be referred to Section 3.1. We denote the image representations of S Net as $\Phi = \{\phi_M^l, l \in [1, 2, 3, 4]\}$, where ϕ_M^i is the ith layer representation for input data M. It maps an input image from modality M to a feature vector in $\mathbb{R}^d$. Specifically, we define the $L1$ loss between two feature vectors for translation supervision. Suppose we are training the classification task for $\mathbf{M_A}$. Generated images y_i' and $\mathbf{M_B}$ are fed into the S Net, and we can get layer-wise presentations from y_i' and $\mathbf{M_B}$. We constrain them from every each layer (layer1-layer4 in ResNet) be the same by $L1$ loss:

$$L_{content}(y_i, y_i', l) = \sum_{l=1}^{L} \|\phi_{y_i}^l - \phi_{T(x_i)}^l\|_1 \qquad (1)$$

3.3. Training Strategy

In this section, we introduce our optimization procedure in detail. To jointly learn the embedding and the translation pair, we optimize E, C, D networks in a multi-task manner. Specifically, given a pair of RGB-D images, let $e_{\mathbf{M_A}} = E(\{x_i\})$ be the embeddings from E Net computed on $\mathbf{M_A}$ and $d_{\mathbf{M_B}} = D(E(\{x_i\})$ be the generated modality B data decoded from D. We simultaneously update 1) the $E \rightarrow D \rightarrow S$ to minimize the distance of layer-wise vectors to constrain their semantic similarity, and 2) the $E \rightarrow C$ using a cross entropy loss function for classification task. The total loss is updated with a linear combination:

$$L_{total} = \alpha L_{content} + \beta L_{C_{cls}} \qquad (2)$$

where $L_{content}$ is Equation (1) described in Section 3.2, L_{cls} is the cross-entropy loss for classification task, with their coefficients α, β set as 10, 1 from the best trials.

As the datasets we use in Section 4 are characterized by unbalanced images for categories, we use rescaling weights given to each category for cross-entropy loss. Rescaling weights aims to assign different weights to different classes to handle the issue of imbalance training. Specifically, we use the following rescaling strategy:

$$L_{weighted\_cls} = \frac{1}{N} \sum_i -w(y_i) \log \frac{f(x_i)_{y_i}}{\sum_j f(x_i)_j}. \qquad (3)$$

The weight $w(y)$ is computed as:

$$w(y) = \frac{N(y) - N_{(c\_min)} + \delta}{N_{(c\_max)} - N_{(c\_min)}}, \qquad (4)$$

where N_y is the number of images of class y. c_min and c_max represent the class with the least and the most number of training images. The δ is set as 0.01.

In the test phase, we only use the recognition branch for recognition prediction, as illustrated in Figure 1.

Imbalanced translations between two modalities. There are several factors demonstrating that the translations from RGB-to-depth and depth-to-RGB are imbalanced. For example, translation from RGB to depth images is more relatively natural procedure while it would become an ill-posed problem from the reversed direction. What's more, the ground truth of depth data comes with much more value errors due to the characteristics of collecting equipments and process. Therefore, we sample a random noise vector from N(0,1) concatenated to the input feature to D in
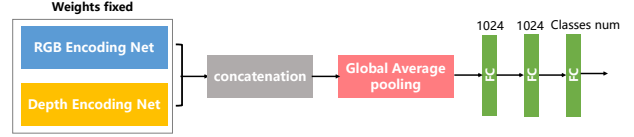

Figure 3. Fusion Network. We only use Encoding Nets from RGB and Depth TRecgNets for fusion. The weights of two E Nets are fixed, and only weights of the classifier are updated.

the Depth TRecgNet (translation from depth to RGB). We found it useful to stable the training of Depth TRecgNet. The generated RGB images also get an interpretable control from the sampled noise. The dimensionality of noise is set as 128 in our experiments.

Initialization with unlabeled RGB-D data. As mentioned in Section 1, the sizes of most existing labeled RGB-D datasets are in small orders of magnitude compared with RGB datasets. However, there are a large number of unlabeled RGB-D pairs, for example, from RGB-D video sequences. A significant advantage of our method is that we are capable of initializing the TRecgNet with these unlabeled data. In other words, the modality translation process is a label-free procedure, by which the TRecgNet could learn rich representations from unlabeled RGB-D data boosting the further task. Related experiments are detailed in Section. 4.

3.4. Fusion

After we get two trained TRecgNets for RGB and depth data, we compute two E Nets and concatenate modality-specific features from them. The embedding is operated on global average pooling (GAP) [18] to reduce the number of parameters followed by three fully connected layers. The whole structure is illustrated in Figure 3. We fix the Encoding Networks and directly train the classifier in an end-to-end manner. We find this would be superior to that directly combines the two prediction results and show the effectiveness of modality-specific networks more straightforwardly.

4. Experiments

In this section, we first introduce the evaluation datasets and the implementation details of our proposed approach. Then we discuss the ablation study of TRecgNet on effectiveness and layer contribution of S. We also compare the quality of generated images with other approaches. Finally, we evaluate the performance of our approach with state-of-the-art methods. We quantitatively report the average accuracy over all scene categories following the conventional evaluation scheme.

4.1. Datasets

SUN RGB-D Dataset is currently the largest RGB-D dataset. It contains RGB-D images from NYU depth v2,

Data	Model	Init	Acc (%)
RGB	ResNet18	Places	47.4
	TRecg	Places	49.8
	TRecg Aug	Places	50.6
Depth	ResNet18	random	38.1
	ResNet18	Places	44.5
	TRecg	random	42.2
	TRecg	Places	46.8
	TRecg	random/unlabeled	44.2
	TRecg	Places/unlabeled	47.6
	TRecg Aug	Places	47.9

Table 1. **Ablation study of TRecg-ResNet18 on recognition performance**. The results are reported on the test set of SUN RGB-D (Top-1's mean accuracy %). "Aug" means using generated data in training.

Berkeley B3DO [12], and SUN3D [33] and is compromised of 3,784 Microsoft Kinect v2 images, 3,389 Asus Xtion images, 2,003 Microsoft Kinect v1 images and 1,159 Intel RealSense images. Following the standard experimental settings stated in [25], we only work with 19 major scene categories which contain more than 80 images. As per standard splits, there are in total 4,845 images for training and 4,659 for testing.

NYU Depth Dataset V2 (NYUD2) is a relatively small dataset; only a few of its 27 indoor categories are well presented. Following the standard split in [24], the categories are grouped into ten including nine most common categories and an *other* category representing the rest. Also, we use 795 / 654 images for training / testing following the standard split.

4.2. Implementation Details

The proposed approach is implemented in popular deep learning framework, Pytorch [20], on an NVIDIA TITAN Xp GPU. We train the network using Adam stochastic optimization [14] to learn network parameters, with the batch size set to 40. The RGB-D images are resized to 256×256 and randomly cropped to 224×224. We train the TRecgNet in 70 epochs, and the learning rate is initialized as 2×10^{-4} at the first 20 epochs and linearly decades in the rest of 50. In the test phase, we use a center crop operation on test images. We employ geocentric HHA (Horizontal disparity, Height above ground and Angle with gravity) [6] to encode depth images, which has been shown better performance to capture the scenes structural and geometrical properties of depth data for kinds of vision tasks. In the subsequent experiments, we separately train two kinds of TRecgNets for evaluation. The basic TRecgNet is trained without using generated data while the TrecgNet Aug refers to training a TRecgNet with generated data from the previous corresponding basic TRecgNet. (RGB TRecgNet Aug
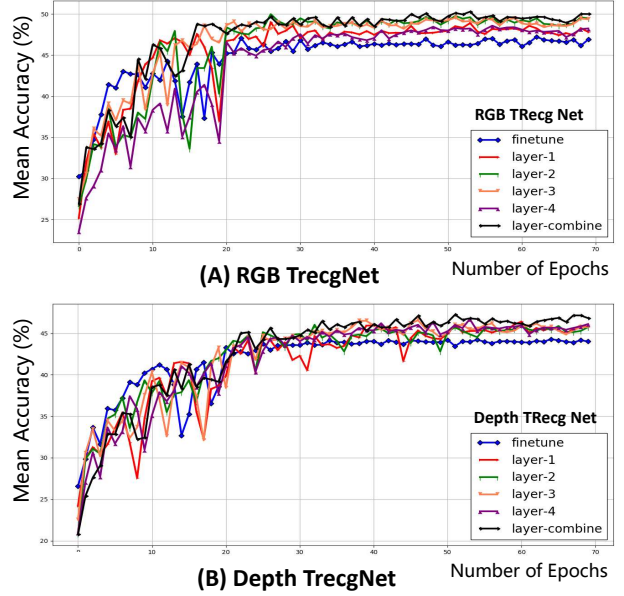

Figure 4. Effect of content model using different layer as supervision on classification for (A) **RGB** TRecgNet and (B) **Depth** TRecgNet. Tested on SUN RGB-D dataset.
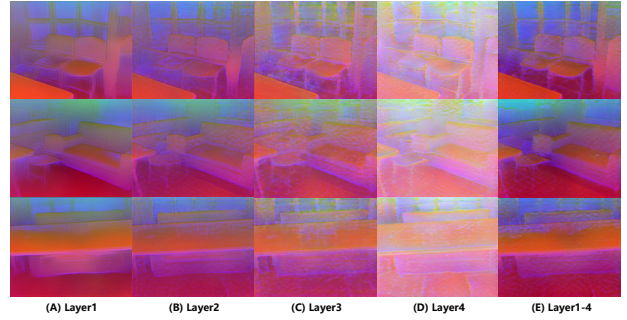

Figure 5. Examples of TRecgNet translating RGB images to depth ones by semantic supervision from different layers. The combination of layer-wise content supervision gives the best photo-realist translation. The input images are all from the test set of SUN RGB-D dataset.

leverages the basic depth TRecgNet as the generated data sampler, and vice versa). Specifically, in the training phase of TrecgNet Aug, we randomly use the generated data, the number of which is controlled as 30% of batch size, to achieve the best performance.

4.3. Results on SUN RGB-D dataset

Study on the effectiveness of TRecgNet. We begin our experiments by studying the effectiveness of TRecgNet for RGB-D recognition task. We tend to prove that 1) learning essential similarity from modality translation branch could effectively benefit classification task, and 2) Unlabeled RGB-D data helps in training TRecgNet for recognition task, and 3) generated data improves training process.

Figure 6. Examples of generated data by our TrecgNet from test set of SUN RGB-D dataset. (B) are depth images translated from original RGB data (A), and (D) are generated RGB images using original depth ones (C).

Loss	Pixel	Pixel+GAN	Ours
Mean Acc (depth)	13.1%	17.7%	30.3%
Mean Acc (rgb)	7.3%	20.6%	18.4%

Table 2. We compare the quality of generated depth images by training a vanilla ResNet18 pre-trained on Places dataset. The training images are all generated using training data of SUN RGB-D dataset by three methods. The results are reported on the test set of SUN RGB-D (Top-1's mean accuracy %). Images generated by our TRecgNet achieve the best result.

Since patterns of color and depth information vary greatly on visual appearance, geometry and surface, we use the following baselines: for the classification of RGB modality, we fine tune pre-trained ResNet18 on RGB images; for the depth modality, we train the network both from scratch and from the pre-trained model.

We test our TRceg Net with following settings: a) the same initialization schedule with baseline, b) the Depth TRecgNet is pre-trained using 5k unlabeled RGB-D video sequences from the NYUD2 dataset. c) Two TRecgNets extract the translation branches as the data augmentation supplier to retrain the TRecgNet; 30% of original training data is randomly replaced with generated data. The experimental results are summarized in Table 1. We observe that for each modality, our TRecgNet outperforms the baseline with big margins. RGB TRecgNet outperforms the baseline by 2.4% while Depth TRecgNet by 3.7% and 2.3% with randomly initialization and pre-trained weights from Places dataset, separately.

It is worth noting that unlabeled RGB-D data pre-training for TRecgNet makes a further boosting of 2% and 0.8%, when training from scratch and using pre-trained weights, respectively. Without pre-trained Places weights but only 5k unlabeled data, the result is very comparable to that of vanilla ResNet pre-trained on Places dataset could achieve, only with quite less training data. This indicates that we could flexibly design a E Net achieving acceptable results without pre-training it using large-scale datasets like ImageNet or Places dataset.

We also find promotions when using generated images as training data, with 1.1% for the Aug-Depth and 0.8% for Aug-RGB TRecgNets. We show some generated data examples in Figure 6. It's interesting to find that the generated depth images tend to be brighter and have better contours than the original ones, due to the learned RGB context information in the translation process. As for generated RGB data, since the original depth data exists with non-negligible measurement error, generated RGB data would be inevitably of low quality on details, however, we turned out to find it be alleviated by adding the random noise in the training of Depth TRecgNet.

Study on layer contribution of Semantic Content Network. The effectiveness of transferring complementary cues relies much on the semantic content model S Net. Therefore, we take an interest in how the S Net affects the translation and recognition tasks using the different layer as supervision. We separately use layer 1 ~ layer 4 as well as their combination as the translation constraint to test the scene recognition performance of TRecgNet, compared with directly fine tuning pre-trained ResNet18 model; all the experiments are conducted on SUN RGB-D dataset. In Figure 4, the accuracy is plotted against the number of training epochs. The following observations can be made: 1) TRecgNet achieves markable improvement to varying degree by using the different layer as supervision. 2) The training procedure becomes more steady when using layer-wise content constraint and achieves the best performance. The result demonstrates the effectiveness of TRecgNet using S as the supervision for cross-modal translation. Figure 5 gives an example of RGB TrecgNet translating images by semantic supervisions from different layers. Similarity learning from lower layers' supervision focus on keeping textures while higher layers' constraints try to capture more

				Accuracy (%)		
	Method	RGB Init	Depth Init	RGB	Depth	Fusion
Baseline	ResNet18	ImageNet	ImageNet	46.6	44.5	50.1
	ResNet18	Places	Places	47.4	44.8	50.8
Proposed	TRecgNet	ImageNet	ImageNet	48.7	46.9	55.5
	TRecgNet Aug	ImageNet	ImageNet	49.2	47.9	**56.1**
	TRecgNet	Places	Places	49.8	46.8	56.1
	TRecgNet Aug	Places	Places	50.6	47.9	**56.7**
SOTA	Multimodal fusion[40]	Places	Places	40.4	36.5	41.5
	Modality&component fusion [27]	Places	Places	40.4	36.5	48.1
	RGB-D-CNN+wSVM [26]	Places	Fast R-CNN	44.6	42.7	53.8
	DF^2Net metric learning [17]	Places	Places	46.3	39.2	54.6

Table 3. Comparison with state-of-the-art methods on the test set of the **SUN RGB-D** dataset. The performance is measured by Top-1's mean accuracy over classes. "Aug" means using generated data in training.

				Accuracy (%)		
	Method	RGB Init	Depth Init	RGB	Depth	Fusion
Baseline	ResNet18	Places	Places	59.8	52.3	63.8
Proposed	TRecgNet	Places	Places	60.2	55.2	65.5
	TRecgNet	SUN RGB-D	SUN RGB-D	63.8	56.7	66.5
	TRecgNet Aug	SUN RGB-D	SUN RGB-D	64.8	57.7	**69.2**
SOTA	Modality & component fusion [27]	Places	Places	53.5	51.5	63.9
	RGB-D-CNN+wSVM [26]	Places	Fast R-CNN	53.4	56.4	67.5
	DF^2Net metric learning [17]	Places	Places	61.1	54.8	65.4

Table 4. Comparison with state-of-the-art methods on the test set of the **NYUD2** dataset. The performance is measured by Top-1's mean accuracy over classes. "Aug" means using generated data in training.

semantic cues ignoring luminance or texture. In contrast, layer-wise content supervision combines multi-layer characteristics which could not only get better performance on classification task but generate more photo-realistic images.

Study on quality of generated images. In this study, we evaluate the quality of generated data from TRecgNet. Since depth data suffers a lot from scarcity and value error problems, we focus on depth data generation. We compare TRecgNet with two image generation approaches. The first one is to generate depth data only using $L1$ loss on pixel-level intensity, which is very common in images reconstruction works [16] as stated before. We also test images generated by pixel-to-pixel GAN supervision [11]. We qualitatively compare some generated examples in Figure 7. Images from pixel intensity supervision tend to be blurred stiffly imitating the training data while our TRecgNet produce more natural images, especially for depth data, even for cases that the ground truth is of significant errors, see the bottom row. Interestingly, RGB images generated based on GAN shows an impressive effect on color diversity. We also quantitatively evaluate the quality of generated images by fine tuning pre-trained ResNet18 only using generated images by different methods. Table 2 shows the results. Depth images from our TRecgNet outperform other methods by a

big margin, indicating its effectiveness for depth data augmentation while GAN based methods achieve a better result on RGB images.

Comparison with state-of-the-art methods. We report TRecgNet on SUN RGB-D test set compared with state-of-the-art methods, as shown in Table 3. Most RGB-D scene recognition methods build upon models pre-trained on Places dataset [39]. Apart from that, we also report the results using ImageNet pre-trained weights. Most of these methods rely on fine tuning Places-CNN. Song *et al.* [26] work on learning more effective depth representations from supervised depth patches via SSP [9] and makes comprehensive investigations on fusion strategies, such as multi-scale and aided from object detection. DF^2Net adopts triplet loss based metric learning [17] to learn discriminative and correlative features for modality-specific representation and fusion learning. Our TRecgNets outperform other state-of-the-art methods on both kinds of modalities and their fusion by an evident margin. It is worth noting that our method succeeds in learning modality-specific features from cross-modal transfer learning, and we do not rely much on any sophisticated and specifically designed fusion strategy.

(A) source modality   **(B) pixel similarity**   **(C) pix2pix GAN**   **(D) ours**   **(E) target modality (GT)**
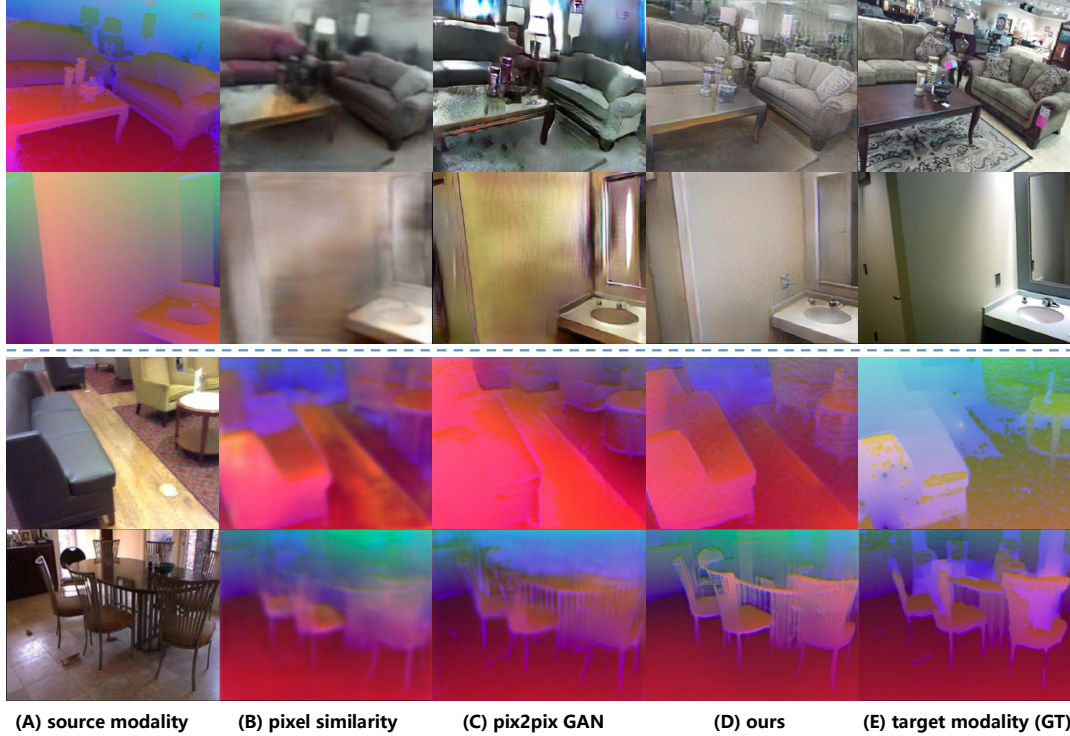
Figure 7. Generated examples by different methods. RGB data translated to depth one is shown in the first two rows while images in the last two rows give a reversed example. (A), (E) are ground truth of RGB and depth data. (B), (C) and (D) are generated images separately from pixel-level L1 loss supervision, pixel-to-pixel GAN, and our TRecgNet.

4.4. Results on NYUD2 dataset

We also evaluate TRecgNet on the NYUD2 test set and compare with other representative works. NYUD2 is a relatively small RGB-D dataset, thus we only evaluate the TRecgNet using pre-trained weights of Places dataset or SUN RGB-D dataset. In particular, we study the generalization ability of learned TRecgNet representations on SUN RGB-D. We transfer the learned TRecgNets from SUN RGB-D and fine tune on data from the NYUD2 dataset. We report the results in Table 4. We find that for RGB modality, our RGB TRecgNet only yields a slightly better result. We argue that it is mainly because the size of NYUD2 dataset is too small. Negative effects from errors of depth maps badly affects the translation from the RGB to the depth. However, transferring pre-trained weights from SUN RGB-D datasets encourages the TRecgNet to overcome this problem, enhance modality-specific representation power and improve its final recognition performance. When adding the generated data, we observe further promotions for both modalities. Experiments in NYUD2 dataset reveal TRecgNet's requirement on the scale of training data to some degree. Translation using small dataset is difficult in benefiting the multi-training task, especially for modalities that exist measuring errors that can't be ignored. Figure 4 also shows some hints for problem that when training TRecgNet in the first several epochs, it tends to behave more unstably than that of fine tuning the backbone network directly and get suboptimal results.

5. Conclusion and Future Work

In this paper, we have presented an effective Translate-to-Recognize Network (TRecgNet) to learn modality-specific RGB-D representation for RGB-D scene recognition task. TRecgNet enables a CNN classification network to learn more discriminative feature by a translation process learning essential similarity with cross-modal data. Training TRecgNet allows using unlabeled RGB-D data as initialization which makes up for the data scarcity problem. As experiments demonstrated on SUN RGB-D and NYUD2 datasets, we both achieve the state-of-the-art results validating the effectiveness of the proposed method. In the future, we plan to try instantiate more as well as deeper CNN models such as ResNet50 and VGG Network. We will also try to handle the big error problem of depth data.

Acknowledgments

References

[1] Dan Banica and Cristian Sminchisescu. Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in rgb-d images. In *CVPR*, pages 3517–3526, 2015.

[2] Yang Chen, Yu-Kun Lai, and Yong-Jin Liu. Cartoongan: Generative adversarial networks for photo cartoonization. In *CVPR*, pages 9465–9474, 2018.

[3] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, pages 2414–2423. IEEE, 2016.

[4] 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*, pages 2672–2680, 2014.

[5] 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*, 112(2):133–149, 2015.

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

[7] Saurabh Gupta, Judy Hoffman, and Jitendra Malik. Cross modal distillation for supervision transfer. In *CVPR*, pages 2827–2836, 2016.

[8] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. In *ICCV*, pages 2980–2988, 2017.

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In *ECCV*, pages 346–361. Springer, 2014.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE, 2017.

[12] Allison Janoch, Sergey Karayev, Yangqing Jia, Jonathan T Barron, Mario Fritz, Kate Saenko, and Trevor Darrell. A category-level 3d object dataset: Putting the kinect to work. In *Consumer Depth Cameras for Computer Vision*, pages 141–165. Springer, 2013.

[13] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012.

[16] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[17] Yabei Li, Junge Zhang, Yanhua Cheng, Kaiqi Huang, and Tieniu Tan. Df2net: Discriminative feature learning and fusion network for rgb-d indoor scene classification. In *AAAI*, 2018.

[18] Min Lin, Qiang Chen, and Shuicheng Yan. Network in network. *arXiv preprint arXiv:1312.4400*, 2013.

[19] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation. In *ICCV*, volume 4, 2017.

[20] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

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

[22] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. In *NIPS*, pages 91–99, 2015.

[23] 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*, pages 234–241. Springer, 2015.

[24] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, pages 746–760. Springer, 2012.

[25] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, pages 567–576. IEEE, 2015.

[26] Xinhang Song, Shuqiang Jiang, Luis Herranz, and Chengpeng Chen. Learning effective rgb-d representations for scene recognition. *IEEE Transactions on Image Processing*, 2018.

[27] Anran Wang, Jianfei Cai, Jiwen Lu, and Tat-Jen Cham. Modality and component aware feature fusion for rgb-d scene classification. In *CVPR*, pages 5995–6004. IEEE, 2016.

[28] Limin Wang, Sheng Guo, Weilin Huang, Yuanjun Xiong, and Yu Qiao. Knowledge guided disambiguation for large-scale scene classification with multi-resolution cnns. *IEEE Trans. Image Processing*, 26(4):2055–2068, 2017.

[29] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *CVPR*, pages 1430–1439, 2018.

[30] Limin Wang, Zhe Wang, Yu Qiao, and Luc Van Gool. Transferring deep object and scene representations for event recognition in still images. *International Journal of Computer Vision*, 126(2-4):390–409, 2018.

[31] Pichao Wang, Wanqing Li, Zhimin Gao, Yuyao Zhang, Chang Tang, and Philip Ogunbona. Scene flow to action map: A new representation for rgb-d based action recognition with convolutional neural networks. In *CVPR*, 2017.

[32] Xiaolong Wang and Abhinav Gupta. Generative image modeling using style and structure adversarial networks. In *ECCV*, pages 318–335. Springer, 2016.

[33] Jianxiong Xiao, Andrew Owens, and Antonio Torralba. Sun3d: A database of big spaces reconstructed using sfm and object labels. In *ICCV*, pages 1625–1632. IEEE, 2013.

[34] Xiangyang Xu, Yuncheng Li, Gangshan Wu, and Jiebo Luo. Multi-modal deep feature learning for rgb-d object detection. *Pattern Recognition*, 72:300–313, 2017.

[35] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, pages 340–349, 2018.

[36] Jing Zhang, Wanqing Li, Philip O Ogunbona, Pichao Wang, and Chang Tang. Rgb-d-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, 2016.

[37] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *European conference on computer vision*, pages 649–666. Springer, 2016.

[38] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 2017.

[39] 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*, pages 487–495, 2014.

[40] Hongyuan Zhu, Jean-Baptiste Weibel, and Shijian Lu. Discriminative multi-modal feature fusion for rgbd indoor scene recognition. In *CVPR*, pages 2969–2976. IEEE, 2016.