

Deep Multimodal Fusion by Channel Exchanging

Yikai Wang¹, Wenbing Huang¹, Fuchun Sun^{1†}, Tingyang Xu², Yu Rong², Junzhou Huang²

¹Beijing National Research Center for Information Science and Technology (BNRist),

State Key Lab on Intelligent Technology and Systems,

Department of Computer Science and Technology, Tsinghua University ²Tencent AI Lab

wangyk17@mails.tsinghua.edu.cn, hwenbing@126.com, fcsun@tsinghua.edu.cn,

tingyangxu@tencent.com, yu.rong@hotmail.com, jzhuang@uta.edu

Abstract

Deep multimodal fusion by using multiple sources of data for classification or regression has exhibited a clear advantage over the unimodal counterpart on various applications. Yet, current methods including aggregation-based and alignment-based fusion are still inadequate in balancing the trade-off between inter-modal fusion and intra-modal processing, incurring a bottleneck of performance improvement. To this end, this paper proposes Channel-Exchanging-Network (CEN), a parameter-free multimodal fusion framework that dynamically exchanges channels between sub-networks of different modalities. Specifically, the channel exchanging process is self-guided by individual channel importance that is measured by the magnitude of Batch-Normalization (BN) scaling factor during training. The validity of such exchanging process is also guaranteed by sharing convolutional filters yet keeping separate BN layers across modalities, which, as an add-on benefit, allows our multimodal architecture to be almost as compact as a unimodal network. Extensive experiments on semantic segmentation via RGB-D data and image translation through multi-domain input verify the effectiveness of our CEN compared to current state-of-the-art methods. Detailed ablation studies have also been carried out, which provably affirm the advantage of each component we propose. Our code is available at <https://github.com/yikaiw/CEN>.

1 Introduction

Encouraged by the growing availability of low-cost sensors, *multimodal fusion* that takes advantage of data obtained from different sources/structures for classification or regression has become a central problem in machine learning [4]. Joining the success of deep learning, multimodal fusion is recently specified as *deep multimodal fusion* by introducing end-to-end neural integration of multiple modalities [38], and it has exhibited remarkable benefits against the unimodal paradigm in semantic segmentation [29, 45], action recognition [14, 15, 44], visual question answering [1, 23], and many others [3, 26, 52].

A variety of works have been done towards deep multimodal fusion [38]. Regarding the type of how they fuse, existing methods are generally categorized into *aggregation-based* fusion, *alignment-based* fusion, and the mixture of them [4]. The aggregation-based methods employ a certain operation (e.g. averaging [19], concatenation [35, 51], and self-attention [45]) to combine multimodal sub-networks into a single network. The alignment-based fusion [9, 44, 47], instead, adopts a regulation loss to align the embedding of all sub-networks while keeping full propagation for each of them. The difference between such two mechanisms is depicted in Figure 1. Another categorization of multimodal fusion can be specified as early, middle, and late fusion, depending on when to fuse, which have been discussed in earlier works [2, 7, 18, 42] and also in the deep learning literature [4, 27, 28, 46].

[†]Corresponding author: Fuchun Sun.

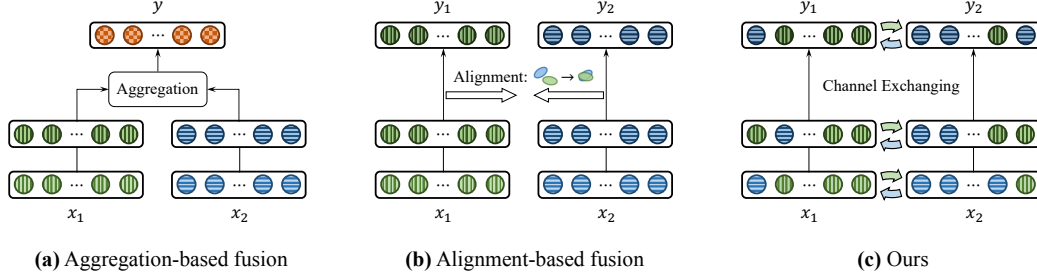


Figure 1: A sketched comparison between existing fusion methods and ours.

Albeit the fruitful progress, it remains a great challenge on how to integrate the common information across modalities, meanwhile preserving the specific patterns of each one. In particular, the aggregation-based fusion is prone to underestimating the intra-modal propagation once the multi-modal sub-networks have been aggregated. On the contrary, the alignment-based fusion maintains the intra-modal propagation, but it always delivers ineffective inter-modal fusion owing to the weak message exchanging by solely training the alignment loss. To balance between inter-modal fusion and intra-modal processing, current methods usually resort to careful hierarchical combination of the aggregation and alignment fusion for enhanced performance, at a cost of extra computation and engineering overhead [12, 29, 51].

Present Work. We propose Channel-Exchanging-Network (CEN) which is parameter-free, adaptive, and effective. Instead of using aggregation or alignment as before, CEN dynamically exchanges the channels between sub-networks for fusion (see Figure 1(c)). The core of CEN lies in its smaller-norm-less-informative assumption inspired from network pruning [33, 49]. To be specific, we utilize the scaling factor (*i.e.* γ) of Batch-Normalization (BN) [24] as the importance measurement of each corresponding channel, and replace the channels associated with close-to-zero factors of each modality with the mean of other modalities. Such message exchanging is parameter-free and self-adaptive, as it is dynamically controlled by the scaling factors that are determined by the training itself. Besides, we only allow directed channel exchanging within a certain range of channels in each modality to preserve intra-modal processing. More details are provided in § 3.3. Necessary theories on the validity of our idea are also presented in § 3.5.

Another hallmark of CEN is that the parameters except BN layers of all sub-networks are shared with each other (§ 3.4). Although this idea is previously studied in [8, 48], we apply it here to serve specific purposes in CEN: by using private BNs, as already discussed above, we can determine the channel importance for each individual modality; by sharing convolutional filters, the corresponding channels among different modalities are embedded with the same mapping, thus more capable of modeling the modality-common statistic. This design further compacts the multimodal architecture to be almost as small as the unimodal one.

We evaluate our CEN on two studies: semantic segmentation via RGB-D data [41, 43] and image translation through multi-domain input [50]. It demonstrates that CEN yields remarkably superior performance than various kinds of fusion methods based on aggregation or alignment under a fair condition of comparison. In terms of semantic segmentation particularly, our CEN significantly outperforms state-of-the-art methods on two popular benchmarks. We also conduct ablation studies to isolate the benefit of each proposed component. More specifications are provided in § 4.

2 Related Work

We introduce the methods of deep multimodal fusion, and the concepts related to our paper.

Deep multimodal fusion. As discussed in introduction, deep multimodal fusion methods can be mainly categorized into aggregation-based fusion and alignment-based fusion [4]. Due to the weakness in intra-modal processing, recent aggregation-based works perform feature fusion while still maintaining the sub-networks of all modalities [12, 30]. Besides, [19] points out the performance by fusion is highly affected by the choice of which layer to fuse. Alignment-based fusion methods align multimodal features by applying the similarity regulation, where Maximum-Mean-Discrepancy

(MMD) [16] is usually adopted for the measurement. However, simply focusing on unifying the whole distribution may overlook the specific patterns in each domain/modality [6, 44]. Hence, [47] provides a way that may alleviate this issue, which correlates modality-common features while simultaneously maintaining modality-specific information. There is also a portion of the multimodal learning literature based on modulation [11, 13, 46]. Different from these types of fusion methods, we propose a new fusion method by channel exchanging, which potentially enjoys the guarantee to both sufficient inter-model interactions and intra-modal learning.

Other related concepts. The idea of using BN scaling factor to evaluate the importance of CNN channels has been studied in network pruning [33, 49] and representation learning [40]. Moreover, [33] enforces ℓ_1 norm penalty on the scaling factors and explicitly prunes out filters meeting a sparsity criteria. Here, we apply this idea as an adaptive tool to determine where to exchange and fuse. CBN [46] performs cross-modal message passing by modulating BN of one modality conditional on the other, which is clearly different from our method that directly exchanges channels between different modalities for fusion. ShuffleNet [53] proposes to shuffle a portion of channels among multiple groups for efficient propagation in light-weight networks, which is similar to our idea of exchanging channels for message fusion. Yet, while the motivation of our paper is highly different, the exchanging process is self-determined by the BN scaling factors, instead of the random exchanging in ShuffleNet.

3 Channel Exchanging Networks

In this section, we introduce our CEN, by mainly specifying its two fundamental components: the channel exchanging process and the sub-network sharing mechanism, followed by necessary analyses.

3.1 Problem Definition

Suppose we have the i -th input data of M modalities, $\mathbf{x}^{(i)} = \{\mathbf{x}_m^{(i)} \in \mathbb{R}^{C \times (H \times W)}\}_{m=1}^M$, where C denotes the number of channels, H and W denote the height and width of the feature map². We define N as the batch-size. The goal of deep multimodal fusion is to determine a multi-layer network $f(\mathbf{x}^{(i)})$ (particularly CNN in this paper) whose output $\hat{\mathbf{y}}^{(i)}$ is expected to fit the target $\mathbf{y}^{(i)}$ as much as possible. This can be implemented by minimizing the empirical loss as

$$\min_f \frac{1}{N} \sum_{i=1}^N \mathcal{L}(\hat{\mathbf{y}}^{(i)} = f(\mathbf{x}^{(i)}), \mathbf{y}^{(i)}). \quad (1)$$

We now introduce two typical kinds of instantiations to Equation 1:

I. The aggregation-based fusion first processes each m -th modality with a separate sub-network f_m and then combine all their outputs via an aggregation operation followed by a global mapping. In formal, it computes the output by

$$\hat{\mathbf{y}}^{(i)} = f(\mathbf{x}^{(i)}) = h(\text{Agg}(f_1(\mathbf{x}_1^{(i)}), \dots, f_M(\mathbf{x}_M^{(i)}))), \quad (2)$$

where h is the global network and Agg is the aggregation function. The aggregation can be implemented as averaging [19], concatenation [51], and self-attention [45]. All networks are optimized via minimizing Equation 1.

II. The alignment-based fusion leverages an alignment loss for capturing the inter-modal concordance while keeping the outputs of all sub-networks f_m . Formally, it solves

$$\min_{f_{1:M}} \frac{1}{N} \sum_{i=1}^N \mathcal{L}\left(\sum_{m=1}^M \alpha_m f_m(\mathbf{x}_m^{(i)}), \mathbf{y}^{(i)}\right) + \text{Alig}_{f_{1:M}}(\mathbf{x}^{(i)}), \quad s.t. \sum_{m=1}^M \alpha_m = 1, \quad (3)$$

where the alignment $\text{Alig}_{f_{1:M}}$ is usually specified as Maximum-Mean-Discrepancy (MMD) [16] between certain hidden features of sub-networks, and the final output $\sum_{m=1}^M \alpha_m f_m(\mathbf{x}_m^{(i)})$ is an

²Although our paper is specifically interested in image data, our method is still general to other domains; for example, we can set $H = W = 1$ for vectors.

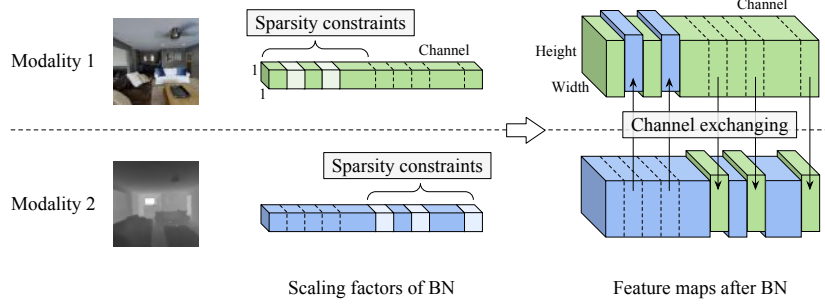


Figure 2: An illustration of our multimodal fusion strategy. The sparsity constraints on scaling factors are applied to disjoint regions of different modalities. A feature map will be replaced by that of other modalities at the same position, if its scaling factor is lower than a threshold.

ensemble of f_m associated with the decision score α_m which is learnt by an additional softmax output to meet the simplex constraint.

As already discussed in introduction, both fusion methods are insufficient to determine the trade-off between fusing modality-common information and preserving modality-specific patterns. In contrast, our CEN is able to combine their best, the details of which are clarified in the next sub-section.

3.2 Overall Framework

The whole optimization objective of our method is

$$\min_{f_{1:M}} \frac{1}{N} \sum_{i=1}^N \mathcal{L} \left(\sum_{m=1}^M \alpha_m f_m(\mathbf{x}^{(i)}), \mathbf{y}^{(i)} \right) + \lambda \sum_{m=1}^M \sum_{l=1}^L |\hat{\gamma}_{m,l}|, \quad s.t. \sum_{m=1}^M \alpha_m = 1, \quad (4)$$

where,

- The sub-network $f_m(\mathbf{x}^{(i)})$ (opposed to $f_m(\mathbf{x}_m^{(i)})$ in Equation 3 of the alignment fusion) fuses multimodal information by channel exchanging, as we will detail in § 3.3;
- Each sub-network is equipped with BN layers containing the scaling factors $\gamma_{m,l}$ for the l -th layer, and we will penalize the ℓ_1 norm of their certain portion $\hat{\gamma}_{m,l}$ for sparsity, which is presented in § 3.3;
- The sub-network f_m shares the same parameters except BN layers to facilitate the channel exchanging as well as to compact the architecture further, as introduced in § 3.4;
- The decision scores of the ensemble output, α_m , are trained by a softmax output similar to the alignment-based methods.

By the design of Equation 4, we conduct a parameter-free message fusion across modalities while maintaining the self-propagation of each sub-network so as to characterize the specific statistic of each modality. Moreover, our fusion of channel exchanging is self-adaptive and easily embedded to everywhere of the sub-networks, with the details given in what follows.

3.3 Channel Exchanging by Comparing BN Scaling Factor

Prior to introducing the channel exchanging process, we first review the BN layer [24], which is used widely in deep learning to eliminate covariate shift and improve generalization. We denote by $\mathbf{x}_{m,l}$ the l -th layer feature maps of the m -th sub-network, and by $\mathbf{x}_{m,l,c}$ the c -th channel. The BN layer performs a normalization of $\mathbf{x}_{m,l}$ followed by an affine transformation, namely,

$$\mathbf{x}'_{m,l,c} = \gamma_{m,l,c} \frac{\mathbf{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}} + \beta_{m,l,c}, \quad (5)$$

where, $\mu_{m,l,c}$ and $\sigma_{m,l,c}$ compute the mean and the standard deviation, respectively, of all activations over all pixel locations (H and W) for the current mini-batch data; $\gamma_{m,l,c}$ and $\beta_{m,l,c}$ are the trainable

scaling factor and offset, respectively; ϵ is a small constant to avoid divisions by zero. The $(l + 1)$ -th layer takes $\{\mathbf{x}'_{m,l,c}\}_c$ as input after a non-linear function.

The factor $\gamma_{m,l,c}$ in Equation 5 evaluates the correlation between the input $\mathbf{x}_{m,l,c}$ and the output $\mathbf{x}'_{m,l,c}$ during training. The gradient of the loss w.r.t. $\mathbf{x}_{m,l,c}$ will approach 0 if $\gamma_{m,l,c} \rightarrow 0$, implying that $\mathbf{x}_{m,l,c}$ will lose its influence to the final prediction and become redundant thereby. Moreover, we will prove in § 3.5 that the state of $\gamma_{m,l,c} = 0$ is attractive with a high probability, given the ℓ_1 norm regulation in Equation 4. In other words, once the current channel $\mathbf{x}_{m,l,c}$ becomes redundant due to $\gamma_{m,l,c} \rightarrow 0$ at a certain training step, it will almost do henceforth.

It thus motivates us to replace the channels of small scaling factors with the ones of other sub-networks, since those channels potentially are redundant. To do so, we derive

$$\mathbf{x}'_{m,l,c} = \begin{cases} \gamma_{m,l,c} \frac{\mathbf{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}} + \beta_{m,l,c}, & \text{if } \gamma_{m,l,c} > \theta; \\ \frac{1}{M-1} \sum_{m' \neq m}^M \gamma_{m',l,c} \frac{\mathbf{x}_{m',l,c} - \mu_{m',l,c}}{\sqrt{\sigma_{m',l,c}^2 + \epsilon}} + \beta_{m',l,c}, & \text{else;} \end{cases} \quad (6)$$

where, the current channel is replaced with the mean of other channels if its scaling factor is smaller than a certain threshold $\theta \approx 0^+$. In a nutshell, if one channel of one modality has little impact to the final prediction, then we replace it with the mean of other modalities. We apply Equation 6 for each modality before feeding them into the nonlinear activation followed by the convolutions in the next layer. Gradients are detached from the replaced channel and back-propagated through the new ones.

In our implementation, we divide the whole channels into M equal sub-parts, and only perform the channel exchanging in each different sub-part for different modality. We denote the scaling factors that are allowed to be replaced as $\hat{\gamma}_{m,l}$. We further impose the sparsity constraint on $\hat{\gamma}_{m,l}$ in Equation 4 to discover unnecessary channels. As the exchanging in Equation 6 is a directed process within only one sub-part of channels, it hopefully can not only retain modal-specific propagation in the other $M - 1$ sub-parts but also avoid unavailing exchanging since $\gamma_{m',l,c}$, different from $\hat{\gamma}_{m,l,c}$, is out of the sparsity constraint. Figure 2 illustrates our channel exchanging process.

3.4 Sub-Network Sharing with Independent BN

It is known in [8, 48] that leveraging private BN layers is able to characterize the traits of different domains or modalities. In our method, specifically, different scaling factors (Equation 5) evaluate the importance of the channels of different modalities, and they should be decoupled.

With the exception of BN layers, all sub-networks f_m share all parameters with each other including convolutional filters³. The hope is that we can further reduce the network complexity and therefore improve the predictive generalization. Rather, considering the specific design of our framework, sharing convolutional filters is able to capture the common patterns in different modalities, which is a crucial purpose of multimodal fusion. In our experiments, we conduct multimodal fusion on RGB-D images or on other domains of images corresponding to the same image content. In this scenario, all modalities are homogeneous in the sense that they are just different views of the same input. Thus, sharing parameters between different sub-networks still yields promisingly expressive power. Nevertheless, when we are dealing with heterogeneous modalities (e.g. images with text sequences), it would impede the expressive power of the sub-networks if keeping sharing their parameters, hence a more dexterous mechanism is suggested, the discussion of which is left for future exploration.

3.5 Analysis

Theorem 1. Suppose $\{\gamma_{m,l,c}\}_{m,l,c}$ are the BN scaling factors of any multimodal fusion network (without channel exchanging) optimized by Equation 4. Then the probability of $\gamma_{m,l,c}$ being attracted to $\gamma_{m,l,c} = 0$ during training (a.k.a. $\gamma_{m,l,c} = 0$ is the local minimum) is equal to $2\Phi(\lambda |\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}}|^{-1}) - 1$, where Φ derives the cumulative probability of standard Gaussian.

In practice, especially when approaching the convergence point, the magnitude of $\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}}$ is usually very close to zero, indicating that the probability of staying around $\gamma_{m,l,c} = 0$ is large. In other words,

³If the input channels of different modalities are different (e.g. RGB and depth), we will broaden their sizes to be the same as their Least Common Multiple (LCM).

Table 1: Detailed results for different versions of our CEN on NYUDv2. All results are obtained with the backbone RefineNet (ResNet101) of single-scale evaluation for test.

Convs	BNs	ℓ_1 Regulation	Exchange	Mean IoU (%)		
				RGB	Depth	Ensemble
Unshared	Unshared	×	×	45.5	35.8	47.6
Shared	Shared	×	×	43.7	35.5	45.2
Shared	Unshared	×	×	46.2	38.4	48.0
Shared	Unshared	Half-channel	×	46.0	38.1	47.7
Shared	Unshared	Half-channel	✓	49.7	45.1	51.1
Shared	Unshared	All-channel	✓	48.6	39.0	49.8

Table 2: Comparison with three typical fusion methods including concatenation (concat), fusion by alignment (align), and self-attention (self-att.) on NYUDv2. All results are obtained with the backbone RefineNet (ResNet101) of single-scale evaluation for test.

Modality	Approach	Commonly-used setting		Same with our setting		Params used for fusion (M)
		Mean IoU (%)	Params in total (M)	Mean IoU (%)	Params in total (M)	
RGB	Uni-modal	45.5	118.1	45.5 / - / -	118.1	-
Depth	Uni-modal	35.8	118.1	- / 35.8 / -	118.1	-
RGB-D	Concat (early)	47.2	120.1	47.0 / 37.5 / 47.6	118.8	0.6
	Concat (middle)	46.7	147.7	46.6 / 37.0 / 47.4	120.3	2.1
	Concat (late)	46.3	169.0	46.3 / 37.2 / 46.9	126.6	8.4
	Concat (all-stage)	47.5	171.7	47.8 / 36.9 / 48.3	129.4	11.2
	Align (early)	46.4	238.8	46.3 / 35.8 / 46.7	120.8	2.6
	Align (middle)	47.9	246.7	47.7 / 36.0 / 48.1	128.7	10.5
	Align (late)	47.6	278.1	47.3 / 35.4 / 47.6	160.1	41.9
	Align (all-stage)	46.8	291.9	46.6 / 35.5 / 47.0	173.9	55.7
	Self-att. (early)	47.8	124.9	47.7 / 38.3 / 48.2	123.6	5.4
	Self-att. (middle)	48.3	166.9	48.0 / 38.1 / 48.7	139.4	21.2
	Self-att. (late)	47.5	245.5	47.6 / 38.1 / 48.3	203.2	84.9
	Self-att. (all-stage)	48.7	272.3	48.5 / 37.7 / 49.1	231.0	112.8
	Ours	-	-	49.7 / 45.1 / 51.1	118.2	0.0

when the scaling factor of one channel is equal to zero, this channel will almost become redundant during later training process, which will be verified by our experiment in the appendix. Therefore, replacing the channels of $\gamma_{m,l,c} = 0$ with other channels (or anything else) will only enhance the trainability of the model. We immediately have the following corollary,

Corollary 1. *If the minimal of Equation 4 implies $\gamma_{m,l,c} = 0$, then the channel exchanging by Equation 6 (assumed no crossmodal parameter sharing) will only decrease the training loss, i.e. $\min_{f'_{1:M}} L \leq \min_{f_{1:M}} L$, given the sufficiently expressive $f'_{1:M}$ and $f_{1:M}$ which denote the cases with and without channel exchanging, respectively.*

4 Experiments

We contrast the performance of CEN against existing multimodal fusion methods on two different tasks: semantic segmentation and image-to-image translation. The frameworks for both tasks are in the encoder-decoder style. Note that we only perform multimodal fusion within the encoders of different modalities throughout the experiments. Our codes are compiled on PyTorch [36].

4.1 Semantic Segmentation

Datasets. We evaluate our method on two public datasets NYUDv2 [41] and SUN RGB-D [43], which consider RGB and depth as input. Regarding NYUDv2, we follow the standard settings and adopt the split of 795 images for training and 654 for testing, with predicting standard 40 classes [17]. SUN RGB-D is one of the most challenging large-scale benchmarks towards indoor semantic segmentation, containing 10,335 RGB-D images of 37 semantic classes. We use the public train-test split (5,285 vs 5,050).

Implementation. We consider RefineNet [32]/PSPNet [54] as our segmentation framework whose backbone is implemented by ResNet [20] pretrained from ImageNet dataset [39]. The initial learn-

Table 3: Comparison with SOTA methods on semantic segmentation.

Modality	Approach	Backbone Network	NYUDv2			SUN RGB-D		
			Pixel Acc. (%)	Mean Acc. (%)	Mean IoU (%)	Pixel Acc. (%)	Mean Acc. (%)	Mean IoU (%)
RGB	FCN-32s [34]	VGG16	60.0	42.2	29.2	68.4	41.1	29.0
	RefineNet [32]	ResNet101	73.8	58.8	46.4	80.8	57.3	46.3
	RefineNet [32]	ResNet152	74.4	59.6	47.6	81.1	57.7	47.0
RGB-D	FuseNet [19]	VGG16	68.1	50.4	37.9	76.3	48.3	37.3
	ACNet [22]	ResNet50	-	-	48.3	-	-	48.1
	SSMA [45]	ResNet50	75.2	60.5	48.7	81.0	58.1	45.7
	SSMA [45] †	ResNet101	75.8	62.3	49.6	81.6	60.4	47.9
	CBN [46] †	ResNet101	75.5	61.2	48.9	81.5	59.8	47.4
	3DGNN [37]	ResNet101	-	-	-	-	57.0	45.9
	SCN [31]	ResNet152	-	-	49.6	-	-	50.7
	CFN [30]	ResNet152	-	-	47.7	-	-	48.1
	RDFNet [29]	ResNet101	75.6	62.2	49.1	80.9	59.6	47.2
	RDFNet [29]	ResNet152	76.0	62.8	50.1	81.5	60.1	47.7
	Ours-RefineNet (single-scale)	ResNet101	76.2	62.8	51.1	82.0	60.9	49.6
	Ours-RefineNet	ResNet101	77.2	63.7	51.7	82.8	61.9	50.2
	Ours-RefineNet	ResNet152	77.4	64.8	52.2	83.2	62.5	50.8
	Ours-PSPNet	ResNet152	77.7	65.0	52.5	83.5	63.2	51.1

† indicates our implemented results.

ing rates are set to 5×10^{-4} and 3×10^{-3} for the encoder and decoder, respectively, both of which are reduced to their halves every 100/150 epochs (total epochs 300/450) on NYUDv2 with ResNet101/ResNet152 and every 20 epochs (total epochs 60) on SUN RGB-D. The mini-batch size, momentum and weight decay are selected as 6, 0.9, and 10^{-5} , respectively, on both datasets. We set $\lambda = 5 \times 10^{-3}$ in Equation 4 and the threshold to $\theta = 2 \times 10^{-2}$ in Equation 6. Unless otherwise specified, we adopt the multi-scale strategy [29, 32] for test. We employ the Mean IoU along with Pixel Accuracy and Mean Accuracy as evaluation metrics following [32]. Full implementation details are referred to our appendix.

The validity of each proposed component. Table 1 summarizes the results of different variants of CEN on NYUDv2. We have the following observations: **1.** Compared to the unshared baseline, sharing the convolutional parameters greatly boosts the performance particularly on the Depth modality (35.8 vs 38.4). Yet, the performance will encounter a clear drop if we additionally share the BN layers. This observation is consistent with our analyses in § 3.4 due to the different roles of convolutional filters and BN parameters. **2.** After carrying out directed channel exchanging under the ℓ_1 regulation, our model gains a huge improvement on both modalities, *i.e.* from 46.0 to 49.7 on RGB, and from 38.1 to 45.1 on Depth, and finally increases the ensemble Mean IoU from 47.6 to 51.1. It thus verifies the effectiveness of our proposed mechanism on this task. **3.** Note that the channel exchanging is only available on a certain portion of each layer (*i.e.* the half of the channels in the two-modal case). When we remove this constraint and allow all channels to be exchanged by Equation 6, the accuracy decreases, which we conjecture is owing to the detriment by impeding modal-specific propagation, if all channels are engaged in cross-modal fusion.

To further explain why channel exchanging works, Figure 3 displays the feature maps of RGB and Depth, where we find that the RGB channel with non-zero scaling factor mainly characterizes the texture, while the Depth channel with non-zero factor focuses more on the boundary; in this sense, performing channel exchanging can better combine the complementary properties of both modalities.

Comparison with other fusion baselines. Table 2 reports the comparison of our CEN with two aggregation-based methods: concatenation [51] and self-attention [45], and one alignment-based approach [47], using the same backbone. All baselines are implemented with the early, middle, late, and all stage fusion. Besides, for a more fair comparison, all baselines are further conducted under the same setting (except channel exchanging) with ours, namely, sharing convolutions with private BNs, and preserving the propagation of all sub-networks. Full details are provided in the appendix. It demonstrates that, on both settings, our method always outperforms others by an average improvement more than 2%. We also report the parameters used for fusion, *e.g.* the aggregation weights of two modalities in concatenation. While self-attention (all-stage) attains the closest performance to us (49.1 vs 51.1), the parameters it used for fusion are considerable, whereas our fusion is parameter-free.

Comparison with SOTAs. We contrast our method against a wide range of state-of-the-art methods. Their results are directly copied from previous papers if provided or re-implemented by us otherwise, with full specifications illustrated in the appendix. Table 3 concludes that our method equipped with PSPNet (ResNet152) achieves new records remarkably superior to previous methods in terms of all

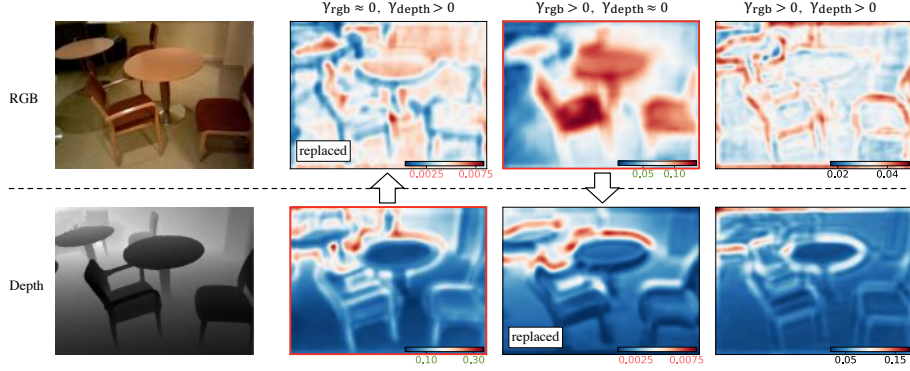


Figure 3: Visualization of the averaged feature maps for RGB and Depth. From left to right: the input images, the channels of $(\gamma_{rgb} \approx 0, \gamma_{depth} > 0)$, $(\gamma_{rgb} > 0, \gamma_{depth} \approx 0)$, and $(\gamma_{rgb} > 0, \gamma_{depth} > 0)$.

Table 4: Comparison on image-to-image translation. Evaluation metrics are FID/KID ($\times 10^{-2}$). Lower values indicate better performance.

Modality	Ours	Baseline	Early	Middle	Late	All-layer
Shade+Texture →RGB	62.63 / 1.65	Concat	87.46 / 3.64	95.16 / 4.67	122.47 / 6.56	78.82 / 3.13
		Average	93.72 / 4.22	93.91 / 4.27	126.74 / 7.10	80.64 / 3.24
		Align	99.68 / 4.93	95.52 / 4.75	98.33 / 4.70	92.30 / 4.20
		Self-att.	83.60 / 3.38	90.79 / 3.92	105.62 / 5.42	73.87 / 2.46
Depth+Normal →RGB	84.33 / 2.70	Concat	105.17 / 5.15	100.29 / 3.37	116.51 / 5.74	99.08 / 4.28
		Average	109.25 / 5.50	104.95 / 4.98	122.42 / 6.76	99.63 / 4.41
		Align	111.65 / 5.53	108.92 / 5.26	105.85 / 4.98	105.03 / 4.91
		Self-att.	100.70 / 4.47	98.63 / 4.35	108.02 / 5.09	96.73 / 3.95

Table 5: Multimodal fusion on image translation (to RGB) with modalities from 1 to 4.

Modality	Depth	Normal	Texture	Shade	Depth+Normal	Depth+Normal +Texture	Depth+Normal +Texture+Shade
FID	113.91	108.20	97.51	100.96	84.33	60.90	57.19
KID ($\times 10^{-2}$)	5.68	5.42	4.82	5.17	2.70	1.56	1.33

metrics on both datasets. In particular, given the same backbone, our method are still much better than RDFNet [29]. To isolate the contribution of RefineNet in our method, Table 3 also provides the uni-modal results, where we observe a clear advantage of multimodal fusion.

Additional ablation studies. In this part, we provide some additional experiments on NYUDv2, with RefineNet (ResNet101). Results are obtained with single-scale evaluation. **1.** As ℓ_1 enables the discovery of unnecessary channels and comes as a pre-condition of Theorem 1, naively exchanging channels with a fixed portion (without using ℓ_1 and threshold) could not reach good performance. For example, exchanging a fixed portion of 30% channels only gets IoU 47.2. We also find by only exchanging 30% channels at each down-sampling stage of the encoder, instead of every 3×3 convolutional layer throughout the encoder (like our CEN), the result becomes 48.6, which is much lower than our CEN (51.1). **2.** In Table 3, we provide results of our implemented CBN [46] by modulating the BN of depth conditional on RGB. The IoUs of CBN with unshared and shared convolutional parameters are 48.3 and 48.9, respectively. **3.** Directly summing activations (discarding the 1st term in Equation 6) results in IoU 48.1, which could reach 48.4 when summing with a learnt soft gate. **4.** If we replace the ensemble of expert with a concat-fusion block, the result will slightly reduce from 51.1 to 50.8. **5.** Besides, we try to exchange channels randomly like ShuffleNet or directly discard unimportant channels without channel exchanging, the IoUs of which are 46.8 and 47.5, respectively. All above ablations support the optimal design of our architecture.

4.2 Image-to-Image Translation

Datasets. We adopt Taskonomy [50], a dataset with 4 million images of indoor scenes of about 600 buildings. Each image in Taskonomy has more than 10 multimodal representations, including depth

(euclidean/zbuffer), shade, normal, texture, edge, principal curvature, etc. For efficiency, we sample 1,000 high-quality multimodal images for training, and 500 for validation.

Implementation. Following Pix2pix [25], we adopt the U-Net-256 structure for image translation with the consistent setups with [25]. The BN computations are replaced with Instance Normalization layers (INs), and our method (Equation 6) is still applicable. We adopt individual INs in the encoder, and share all other parameters including INs in the decoder. We set λ to 10^{-3} for sparsity constraints and the threshold θ to 10^{-2} . We adopt FID [21] and KID [5] as evaluation metrics, which will be introduced in our appendix.

Comparison with other fusion baselines. In Table 4, we evaluate the performance on two specific translation cases, *i.e.* Shade+Texture→RGB and Depth+Normal→RGB, with more examples included in the appendix. In addition to the three baselines used in semantic segmentation (Concat, Self-attention, Align), we conduct an extra aggregation-based method by using the average operation. All baselines perform fusion under 4 different kinds of strategies: early (at the 1st conv-layer), middle (the 4th conv-layer), late (the 8th conv-layer), and all-layer fusion. As shown in Table 4, our method yields much lower FID/KID than others, which supports the benefit of our proposed idea once again.

Considering more modalities. We now test whether our method is applicable to the case with more than 2 modalities. For this purpose, Table 5 presents the results of image translation to RGB by inputting from 1 to 4 modalities of Depth, Normal, Texture, and Shade. It is observed that increasing the number of modalities improves the performance consistently, suggesting much potential of applying our method towards various cases.

5 Conclusion

In this work, we propose Channel-Exchanging-Network (CEN), a novel framework for deep multi-modal fusion, which differs greatly with existing aggregation-based and alignment-based multimodal fusion. The motivation behind is to boost inter-modal fusion while simultaneously keeping sufficient intra-modal processing. The channel exchanging is self-guided by channel importance measured by individual BNs, making our framework self-adaptive and compact. Extensive evaluations verify the effectiveness of our method.

Acknowledgement

This work is jointly funded by National Natural Science Foundation of China and German Research Foundation (NSFC 61621136008/DFG TRR-169) in project “Crossmodal Learning” II, Tencent AI Lab Rhino-Bird Visiting Scholars Program (VS202006), and China Postdoctoral Science Foundation (Grant No.2020M670337).

Broader Impact

This research enables fusing complementary information from different modalities effectively, which helps improve performance for autonomous vehicles and indoor manipulation robots, also making them more robust to environmental conditions, *e.g.* light, weather. Besides, instead of carefully designing hierarchical fusion strategies in existing methods, a global criterion is applied in our work for guiding multimodal fusion, which allows easier model deployment for practical applications. A drawback of bringing deep neural networks into multimodal fusion is its insufficient interpretability.

References

- [1] Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: ICCV (2015)
- [2] Atrey, P.K., Hossain, M.A., El Saddik, A., Kankanhalli, M.S.: Multimodal fusion for multimedia analysis: a survey. In: Multimedia systems (2010)
- [3] Balntas, V., Doumanoglou, A., Sahin, C., Sock, J., Kouskouridas, R., Kim, T.: Pose guided RGBD feature learning for 3d object pose estimation. In: ICCV (2017)
- [4] Baltrusaitis, T., Ahuja, C., Morency, L.: Multimodal machine learning: A survey and taxonomy. In: IEEE Trans. PAMI (2019)

- [5] Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD gans. In: ICLR (2018)
- [6] Bousmalis, K., Trigeorgis, G., Silberman, N., Krishnan, D., Erhan, D.: Domain separation networks. In: NIPS (2016)
- [7] Bruni, E., Tran, N.K., Baroni, M.: Multimodal distributional semantics. In: Journal of Artificial Intelligence Research (2014)
- [8] Chang, W., You, T., Seo, S., Kwak, S., Han, B.: Domain-specific batch normalization for unsupervised domain adaptation. In: CVPR (2019)
- [9] Cheng, Y., Cai, R., Li, Z., Zhao, X., Huang, K.: Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation. In: CVPR (2017)
- [10] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
- [11] De Vries, H., Strub, F., Chandar, S., Pietquin, O., Larochelle, H., Courville, A.: Guesswhat?! visual object discovery through multi-modal dialogue. In: CVPR (2017)
- [12] Du, D., Wang, L., Wang, H., Zhao, K., Wu, G.: Translate-to-recognize networks for RGB-D scene recognition. In: CVPR (2019)
- [13] Dumoulin, V., Perez, E., Schucher, N., Strub, F., Vries, H.d., Courville, A., Bengio, Y.: Feature-wise transformations. In: Distill (2018)
- [14] Fan, L., Huang, W., Gan, C., Ermon, S., Gong, B., Huang, J.: End-to-end learning of motion representation for video understanding. In: CVPR (2018)
- [15] Garcia, N.C., Morerio, P., Murino, V.: Modality distillation with multiple stream networks for action recognition. In: ECCV (2018)
- [16] Gretton, A., Borgwardt, K.M., Rasch, M.J., Schölkopf, B., Smola, A.J.: A kernel two-sample test. In: JMLR (2012)
- [17] Gupta, S., Arbelaez, P., Malik, J.: Perceptual organization and recognition of indoor scenes from RGB-D images. In: CVPR (2013)
- [18] Hall, D.L., Llinas, J.: An introduction to multisensor data fusion. In: Proceedings of the IEEE (1997)
- [19] Hazirbas, C., Ma, L., Domokos, C., Cremers, D.: Fusetnet: Incorporating depth into semantic segmentation via fusion-based CNN architecture. In: ACCV (2016)
- [20] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
- [21] Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: NIPS (2017)
- [22] Hu, X., Yang, K., Fei, L., Wang, K.: ACNET: attention based network to exploit complementary features for RGBD semantic segmentation. In: ICIP (2019)
- [23] Ilievski, I., Feng, J.: Multimodal learning and reasoning for visual question answering. In: NIPS (2017)
- [24] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: ICML (2015)
- [25] Isola, P., Zhu, J., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: CVPR (2017)
- [26] Jin, W., Yang, K., Barzilay, R., Jaakkola, T.S.: Learning multimodal graph-to-graph translation for molecule optimization. In: ICLR (2019)
- [27] Kiela, D.: Deep embodiment: grounding semantics in perceptual modalities. In: Technical Report (2017)
- [28] Lazaridou, A., Bruni, E., Baroni, M.: Is this a wampimuk? cross-modal mapping between distributional semantics and the visual world. In: ACL (2014)
- [29] Lee, S., Park, S., Hong, K.: Rdfnet: RGB-D multi-level residual feature fusion for indoor semantic segmentation. In: ICCV (2017)
- [30] Lin, D., Chen, G., Cohen-Or, D., Heng, P., Huang, H.: Cascaded feature network for semantic segmentation of RGB-D images. In: ICCV (2017)
- [31] Lin, D., Zhang, R., Ji, Y., Li, P., Huang, H.: SCN: switchable context network for semantic segmentation of RGB-D images. In: IEEE Trans. Cybern. (2020)
- [32] Lin, G., Liu, F., Milan, A., Shen, C., Reid, I.: Refinenet: Multi-path refinement networks for dense prediction. In: IEEE Trans. PAMI (2019)
- [33] Liu, Z., Li, J., Shen, Z., Huang, G., Yan, S., Zhang, C.: Learning efficient convolutional networks through network slimming. In: ICCV (2017)

- [34] Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
- [35] Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., Ng, A.Y.: Multimodal deep learning. In: ICML (2011)
- [36] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: NeurIPS (2019)
- [37] Qi, X., Liao, R., Jia, J., Fidler, S., Urtasun, R.: 3d graph neural networks for RGBD semantic segmentation. In: ICCV (2017)
- [38] Ramachandram, D., Taylor, G.W.: Deep multimodal learning: A survey on recent advances and trends. In: IEEE Signal Processing Magazine (2017)
- [39] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M.S., Berg, A.C., Li, F.: Imagenet large scale visual recognition challenge. In: IJCV (2015)
- [40] Shao, W., Tang, S., Pan, X., Tan, P., Wang, X., Luo, P.: Channel equilibrium networks for learning deep representation. In: ICML (2020)
- [41] Silberman, N., Hoiem, D., Kohli, P., Fergus, R.: Indoor segmentation and support inference from RGBD images. In: ECCV (2012)
- [42] Snoek, C.G., Worring, M., Smeulders, A.W.: Early versus late fusion in semantic video analysis. In: ACM MM (2005)
- [43] Song, S., Lichtenberg, S.P., Xiao, J.: SUN RGB-D: A RGB-D scene understanding benchmark suite. In: CVPR (2015)
- [44] Song, S., Liu, J., Li, Y., Guo, Z.: Modality compensation network: Cross-modal adaptation for action recognition. In: IEEE Trans. Image Process. (2020)
- [45] Valada, A., Mohan, R., Burgard, W.: Self-supervised model adaptation for multimodal semantic segmentation. In: IJCV (2020)
- [46] de Vries, H., Strub, F., Mary, J., Larochelle, H., Pietquin, O., Courville, A.C.: Modulating early visual processing by language. In: NIPS (2017)
- [47] Wang, J., Wang, Z., Tao, D., See, S., Wang, G.: Learning common and specific features for RGB-D semantic segmentation with deconvolutional networks. In: ECCV (2016)
- [48] Wang, Y., Sun, F., Lu, M., Yao, A.: Learning deep multimodal feature representation with asymmetric multi-layer fusion. In: ACM MM (2020)
- [49] Ye, J., Lu, X., Lin, Z., Wang, J.Z.: Rethinking the smaller-norm-less-informative assumption in channel pruning of convolution layers. In: ICLR (2018)
- [50] Zamir, A.R., Sax, A., Shen, W.B., Guibas, L.J., Malik, J., Savarese, S.: Taskonomy: Disentangling task transfer learning. In: CVPR (2018)
- [51] Zeng, J., Tong, Y., Huang, Y., Yan, Q., Sun, W., Chen, J., Wang, Y.: Deep surface normal estimation with hierarchical RGB-D fusion. In: CVPR (2019)
- [52] Zhang, W., Zhou, H., Sun, S., Wang, Z., Shi, J., Loy, C.C.: Robust multi-modality multi-object tracking. In: ICCV (2019)
- [53] Zhang, X., Zhou, X., Lin, M., Sun, J.: Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: CVPR (2018)
- [54] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)

Appendix

A Proofs

Theorem 1. Suppose $\{\gamma_{m,l,c}\}_{m,l,c}$ are the BN scaling factors of any multimodal fusion network (without channel exchanging) optimized by Equation 4. Then the probability of $\gamma_{m,l,c}$ being attracted to $\gamma_{m,l,c} = 0$ during training (a.k.a. $\gamma_{m,l,c} = 0$ is the local minimum) is equal to $2\Phi(\lambda|\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}}|^{-1}) - 1$, where Φ derives the cumulative probability of standard Gaussian.

Proof. The proof is straightforward, since the gradient of L w.r.t. $\gamma_{m,l,c}$ is $\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}} \frac{\mathbf{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}} + \lambda$ when $\gamma_{m,l,c} > 0$, or $\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}} \frac{\mathbf{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}} - \lambda$ when $\gamma_{m,l,c} < 0$ ⁴, according to the BN definition in Equation 5 and the ℓ_1 norm in Equation 4. Staying around $\gamma_{m,l,c} = 0$ during training implies that $\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}} \frac{\mathbf{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}} + \lambda > 0$ as well as $\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}} \frac{\mathbf{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}} - \lambda < 0$, the probability of which is $2\Phi(\lambda|\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}}|^{-1}) - 1$ given that the quantity $\frac{\mathbf{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}}$ can be considered as a random variable of standard Gaussian according to the central limit theorem. \square

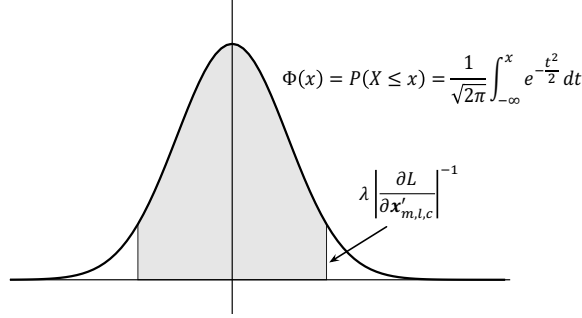


Figure 4: Illustration of the conclusion by Theorem 1.

Corollary 1. If the minimal of Equation 4 implies $\gamma_{m,l,c} = 0$, then the channel exchanging by Equation 6 (assumed no crossmodal parameter sharing) will only decrease the training loss, i.e. $\min_{f'_{1:M}} L \leq \min_{f_{1:M}} L$, given the sufficiently expressive $f'_{1:M}$ and $f_{1:M}$ which denote the cases with and without channel exchanging, respectively.

Proof. We only need to prove for any $f_{1:M}$, we can design a specific $f'_{1:M}$ that shares the same output as $f_{1:M}$ if $\gamma_{m,l,c} = 0$.

- In $f_{1:M}$, the BN layer is followed by a ReLU function and a convolutional layer. We suppose the following convolutional weight for the c -th input channel $\mathbf{x}'_{m,l,c}$ is $\mathbf{W}_{m,l+1,c}$ and the bias is $b_{m,l+1}$. Thus, the quantity related to $\mathbf{x}'_{m,l,c}$ in the $(l+1)$ -th layer is $\mathbf{W}_{m,l+1,c} \otimes \sigma(\mathbf{x}'_{m,l,c}) + b_{m,l+1}$, where \otimes denotes the convolution operation and σ is the ReLU function. Since $\gamma_{m,l,c} = 0$, this term can be translated as $\mathbf{W}_{m,l+1,c} \otimes \sigma(\beta_{m,l,c}) + b_{m,l+1}$, which is a constant feature map.
- As for $f'_{1:M}$, we apply the similar denotations, and attain the term related to $\mathbf{x}'_{m,l,c}$ in the $(l+1)$ -th layer as $\mathbf{W}'_{m,l+1,c} \otimes \sigma(\mathbf{x}'_{m,l,c}) + b'_{m,l+1}$.

By setting $b'_{m,l+1} = \mathbf{W}_{m,l+1,c} \otimes \sigma(\beta_{m,l,c}) + b_{m,l+1}$ and $\mathbf{W}'_{m,l+1,c} = 0$, we will always have $f'_{1:M} = f_{1:M}$, which concludes the proof. \square

⁴Here, we denote $\frac{\partial L}{\partial \mathbf{x}'_{m,l,c}} \frac{\mathbf{x}_{m,l,c} - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}} = \sum_{(i,j)=1}^{(H,W)} \frac{\partial L}{\partial \mathbf{x}'_{m,l,c}}(i,j) \frac{\mathbf{x}_{m,l,c}(i,j) - \mu_{m,l,c}}{\sqrt{\sigma_{m,l,c}^2 + \epsilon}}$ for alleviation, where i, j range over each pixel in $\mathbf{x}'_{m,l,c}$ or $\mathbf{x}_{m,l,c}$.

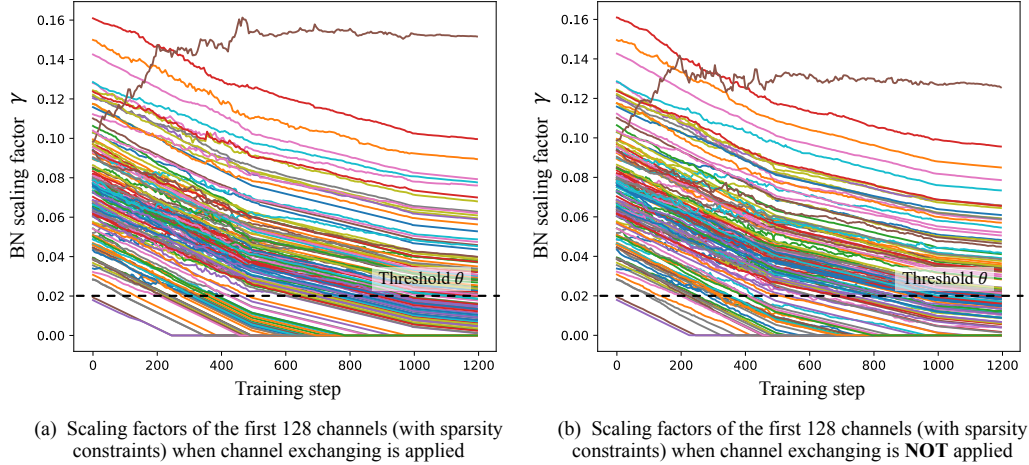


Figure 5: We plot BN scaling factors with sparsity constraints vs training steps. We observe that whether using channel exchanging or not, γ that closes to zero can hardly recover, which verifies our conjecture in Theorem 1. The experiment is conducted on NYUDv2 with RefineNet (ResNet101). We choose the 8th layer of convolutional layers that have 3×3 kernels, and there are totally 256 channels in this layer. Regarding the RGB modality, the sparsity constraints to BN scaling factors are applied for the first 128 channels.

In Figure 4, we provide an illustration of the conclusion by Theorem 1. In Figure 5, we provide experimental results to verify our conjecture in Theorem 1, *i.e.* when the scaling factor of one channel is equal to zero at a certain training step, this channel will almost become redundant during later training process.

In summary, we know that ℓ_1 makes the parameters sparse, but it can not tell if each sparse parameter will keep small in training considering the gradient in Equation 4. Conditional on BN, Theorem 1 proves that $\gamma = 0$ is attractive. Corollary 1 states that f' is more expressive than f when $\gamma = 0$, and thus the optimal f' always outputs no higher loss, which, yet, is not true for arbitrary f' (*e.g.* $f' = 10^6$). Besides, as stated, Corollary 1 holds upon unshared convolutional parameters, and is consistent with Table 7 in the unshared scenario (full-channel: 49.1 vs half-channel: 48.5), although full-channel exchanging is worse under the sharing setting.

B Implementation Details

In our experiments, we adopt ResNet101, ResNet152 for semantic segmentation and U-Net-256 for image-to-image translation. Regarding both ResNet structures, we apply sparsity constraints on Batch-Normalization (BN) scaling factors *w.r.t.* each convolutional layer (conv) with 3×3 kernels. These scaling factors further guide the channel exchanging process that exchanges a portion of feature maps after BN. For the conv layer with 7×7 kernels at the beginning of ResNet, and all other conv layers with 1×1 kernels, we do not apply sparsity constraints or channel exchanging. For U-Net, we apply sparsity constraints on Instance-Normalization (IN) scaling factors *w.r.t.* all conv layers (eight layers in total) in the encoder of the generator, and each is followed by channel exchanging.

We mainly use three multimodal fusion baselines in our paper, including concatenation, alignment and self-attention. Regarding the concatenation method, we stack multimodal feature maps along the channel, and then add a 1×1 convolutional layer to reduce the number of channels back to the original number. The alignment fusion method is a re-implementation of [47], and we follow its default settings for hyper-parameter, *e.g.* using 11 kernel functions for the multiple kernel Maximum Mean Discrepancy. The self-attention method is a re-implementation of the SSMA block proposed in [45], where we also follow the default settings, *e.g.* setting the channel reduction ratio η to 16.

In Table 2, we adopt early, middle, late and all-stage fusion for each baseline method. In ResNet101, there are four stages with 3, 4, 23, 3 blocks, respectively. The early fusion, middle fusion and late

fusion refer to fusing after the 2nd stage, 3rd stage and 4th stage respectively. All-stage fusion refers to fusing after the four stages.

We use a NVIDIA Tesla V100 with 32GB for the experiments.

We now introduce the metrics used in our image-to-image translation task. In Table 4, we adopt the following evaluation metrics:

Fréchet-Inception-Distance (FID) proposed by [21], contrasts the statistics of generated samples against real samples. The FID fits a Gaussian distribution to the hidden activations of InceptionNet for each compared image set and then computes the Fréchet distance (also known as the Wasserstein-2 distance) between those Gaussians. Lower FID is better, corresponding to generated images more similar to the real.

Kernel-Inception-Distance (KID) developed by [5], is a metric similar to the FID but uses the squared Maximum-Mean-Discrepancy (MMD) between Inception representations with a polynomial kernel. Unlike FID, KID has a simple unbiased estimator, making it more reliable especially when there are much more inception features channels than image numbers. Lower KID indicates more visual similarity between real and generated images. Regarding our implementation of KID, the hidden representations are derived from the Inception-v3 pool3 layer.

C Additional Results

We provide three more image translation cases in Table 6, including RGB+Shade→Normal, RGB+Normal→Shade and RGB+Edge→Depth. For baseline methods, we adopt the same settings with Table 4, by adopting early (at the 1st conv-layer), middle (the 4th conv-layer), late (the 8th conv-layer) and all-layer fusion. We adopt MAE (L1 loss) and MSE (L2 loss) as evaluation metrics, and lower values indicate better performance. Our method yields lower MAE and MSE than baseline methods.

Table 6: Comparison on image-to-image translation. Evaluation metrics adopted are MAE ($\times 10^{-1}$)/MSE ($\times 10^{-1}$). Lower values indicate better performance.

Modality	Ours	Baseline	Early	Middle	Late	All-layer
RGB+Shade →Normal	1.12 / 2.51	Concat	1.33 / 2.83	1.22 / 2.65	1.39 / 2.88	1.34 / 2.85
		Average	1.42 / 3.05	1.26 / 2.70	1.40 / 2.90	1.28 / 2.83
		Align	1.45 / 3.11	1.39 / 2.93	1.28 / 2.76	1.52 / 3.25
		Self-att.	1.30 / 2.82	1.18 / 2.59	1.42 / 2.91	1.26 / 2.76
RGB+Normal →Shade	1.10 / 1.72	Concat	1.56 / 2.45	1.38 / 2.12	1.26 / 1.92	1.28 / 2.02
		Average	1.46 / 2.29	1.28 / 2.04	1.51 / 2.39	1.23 / 1.86
		Align	1.39 / 2.26	1.32 / 2.16	1.27 / 2.04	1.41 / 2.21
		Self-att.	1.21 / 1.83	1.15 / 1.73	1.45 / 2.28	1.18 / 1.76
RGB+Edge →Depth	0.28 / 0.66	Concat	0.34 / 0.75	0.32 / 0.74	0.38 / 0.79	0.33 / 0.75
		Average	0.36 / 0.78	0.34 / 0.76	0.36 / 0.77	0.33 / 0.74
		Align	0.44 / 0.89	0.39 / 0.82	0.42 / 0.86	0.44 / 0.90
		Self-att.	0.30 / 0.71	0.33 / 0.73	0.34 / 0.75	0.30 / 0.70

D Results Visualization

In Figure 6 and Figure 7, we provide results visualization for the semantic segmentation task. We choose three baselines including concatenation (concat), alignment (align) and self-attention (self-att.). Among them, concatenation and self-attention methods adopt all-stage fusion, and the alignment method adopts middle fusion (fusion at the end of the 2nd ResNet stage).

In Figure 8, Figure 9 and Figure 10, we provide results visualization for the image translation task. Regarding this task, concatenation and self-attention methods adopt all-layer fusion (fusion at all eight layers in the encoder), and the alignment method adopts middle fusion (fusion at the 4th layer). We adopt these settings in order to achieve high performance for each baseline method.

In the captions of these figures, we detail the prediction difference of different methods.

E Ablation Studies

In Table 7, we provide more cases as a supplement to Table 1. Specifically, we compare the results of channel exchanging when using shared/unshared conv parameters. According to these results, we believe our method is generally useful and channels are aligned to some extent even under the unshared setting.

In Table 8, we verify that sharing convolutional layers (convs) but using individual Instance-Normalization layers (INs) allows 2~4 modalities trained in a single network, achieving even better performance than training with individual networks. Again, if we further sharing INs, there will be an obvious performance drop. More detailed comparison is provided in Table 9.

For the experiment Shade+Texture+Depth→RGB with shared convs and unshared INs, in Figure 11, we plot the proportion of IN scaling factors at the 7th conv layer in the encoder of U-Net. We compare the scaling factors when no sparsity constraints, sparsity constraints applied on all channels, and sparsity constraints applied on disjoint channels. In Figure 12, we further compare scaling factors on all conv layers. In Figure 13, we provide sensitivity analysis for λ and θ .

Table 7: Supplement to Table 1 with more cases. Detailed results for different versions of our CEN on NYUDv2. All results are obtained with the backbone RefineNet (ResNet101) of single-scale evaluation for test. We observe that sharing convs (with unshared BNs) results in better performance for our method.

Convs	BNs	ℓ_1 Regulation	Exchange	Mean IoU (%)		
				RGB	Depth	Ensemble
Unshared	Unshared	×	×	45.5	35.8	47.6
Shared	Shared	×	×	43.7	35.5	45.2
Shared	Unshared	×	×	46.2	38.4	48.0
Unshared	Unshared	Half-channel	×	45.1	35.5	47.3
Unshared	Unshared	Half-channel	✓	46.5	41.6	48.5
Shared	Unshared	Half-channel	×	46.0	38.1	47.7
Shared	Unshared	Half-channel	✓	49.7	45.1	51.1
Unshared	Unshared	All-channel	×	44.6	35.3	46.6
Unshared	Unshared	All-channel	✓	46.8	41.7	49.1
Shared	Unshared	All-channel	×	46.1	37.9	47.5
Shared	Unshared	All-channel	✓	48.6	39.0	49.8

Table 8: We compare training multimodal features in a parallel manner with different parameter sharing settings. Results of the proposed fusion method are reported at the last column. Evaluation metrics are FID/KID ($\times 10^{-2}$). We observe that the convolutional layers can be shared as long as we leave individual INs for different modalities, achieving even better performance.

Modality	Network stream	Unshared convs unshared INs	Shared convs shared INs	Shared convs unshared INs	Multi-modal fusion
Shade+Texture →RGB	Shade	102.21 / 5.25	112.40 / 5.58	100.69 / 4.51	72.07 / 2.32
	Texture	98.19 / 4.83	102.28 / 5.22	93.40 / 4.18	65.60 / 1.82
	Ensemble	92.72 / 4.15	96.31 / 4.36	87.91 / 3.73	62.63 / 1.65
Shade+Texture +Depth →RGB	Shade	101.86 / 5.18	115.51 / 5.77	98.49 / 4.07	69.37 / 2.21
	Texture	98.60 / 4.89	104.39 / 4.54	95.87 / 4.27	64.70 / 1.73
	Depth	114.18 / 5.71	121.40 / 6.23	107.07 / 5.19	71.61 / 2.27
	Ensemble	91.30 / 3.92	100.41 / 4.73	84.39 / 3.45	58.35 / 1.42
Shade+Texture +Depth+Normal →RGB	Shade	100.83 / 5.06	131.74 / 7.48	96.98 / 4.23	68.70 / 2.14
	Texture	97.34 / 4.77	109.45 / 4.86	94.64 / 4.22	63.26 / 1.69
	Depth	114.50 / 5.83	125.54 / 6.48	109.93 / 5.41	70.47 / 2.09
	Normal	108.65 / 5.45	113.15 / 5.72	99.38 / 4.45	67.73 / 1.98
	Ensemble	89.52 / 3.80	102.78 / 4.67	86.76 / 3.63	57.19 / 1.33

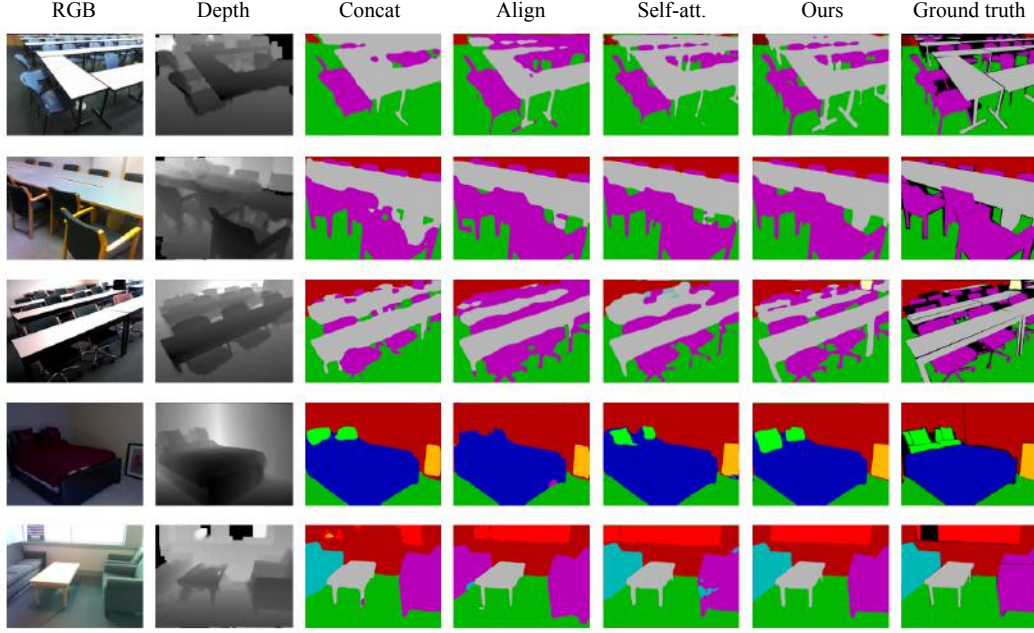


Figure 6: Visualization results for semantic segmentation. Images are collected from NYUDv2 and SUN RGB-D dataset. All results are obtained with the backbone RefineNet (ResNet101) of single-scale evaluation for test. We choose tough images where a number of tables and chairs need to be predicted. Besides, we compare segmentation results on images with low/high light intensity. we observe that the concatenation method is more sensitive to noises of the depth input (see the window at bottom line). Both concatenation and self-attention methods are weak in predicting thin objects *e.g.* table legs and chair legs. These objects are usually missed in the depth input, which may disturb the prediction results during fusion. Compared to baseline fusion methods, the prediction results of our method preserve more details, and are more robust to the light intensity.

Table 9: An Instance-Normalization layer consists of four components, including scaling factors γ , offsets β , running mean μ and variance σ^2 . Following Table 5, we further compare the evaluation results when using unshared γ, β only, and using unshared μ, σ^2 only. Evaluation metrics are FID/KID ($\times 10^{-2}$). We observe these four components of INs are all essential to be unshared. Besides, using unshared scaling factors and offsets seems to be more important.

Modality	Network stream	Unshared convs unshared INs	Shared convs unshared INs	Shared convs, γ, β unshared μ, σ^2	Shared convs, μ, σ^2 unshared γ, β
Shade+Texture +Depth →RGB	Shade	101.86 / 5.18	98.49 / 4.07	107.86 / 5.53	105.29 / 5.29
	Texture	98.60 / 4.89	95.87 / 4.27	105.46 / 5.25	102.90 / 5.06
	Depth	114.18 / 5.71	102.07 / 4.89	118.35 / 6.07	114.35 / 5.80
	Ensemble	91.30 / 3.92	84.39 / 3.45	96.30 / 4.41	92.25 / 4.02
Shade+Texture +Depth+Normal →RGB	Shade	100.83 / 5.06	96.98 / 4.23	113.56 / 5.65	102.74 / 5.17
	Texture	97.34 / 4.77	94.64 / 4.22	105.36 / 5.32	97.53 / 4.56
	Depth	114.50 / 5.83	109.93 / 5.41	119.31 / 6.20	112.73 / 5.60
	Normal	108.65 / 5.45	99.38 / 4.45	108.01 / 5.06	100.34 / 4.53
	Ensemble	89.52 / 3.80	86.76 / 3.63	95.56 / 4.64	89.26 / 3.91

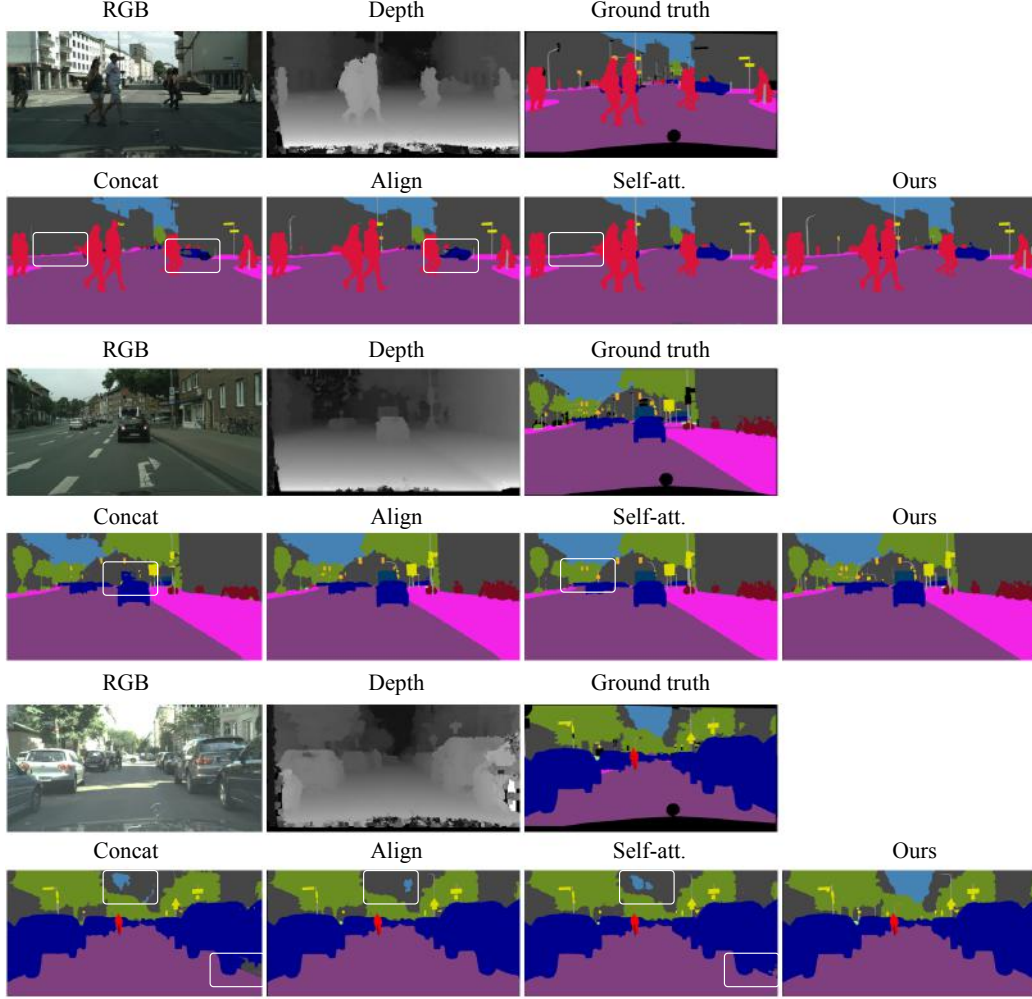


Figure 7: Visualization results for semantic segmentation on Cityscapes dataset [10]. All results are obtained with the backbone PSPNet (ResNet101) of single-scale evaluation for test. Cityscapes is an outdoor dataset containing images from 27 cities in Germany and neighboring countries. The dataset contains 2,975 training, 500 validation and 1,525 test images. There are 20,000 additional coarse annotations provided by the dataset, which are not used for training in our experiments. For the baseline methods, we use white frames to highlight the regions with poor prediction results. We can observe that when the light intensity is high, the baseline methods are weak in capturing the boundary between the sky and buildings using the depth information. Besides, the concatenation and self-attention methods do not preserve fine-grained objects, *e.g.* traffic signs, and are sensitive to noises of the depth input (see the rightmost vehicle in the first group). In contrast, the prediction of our method are better at these aforementioned aspects.

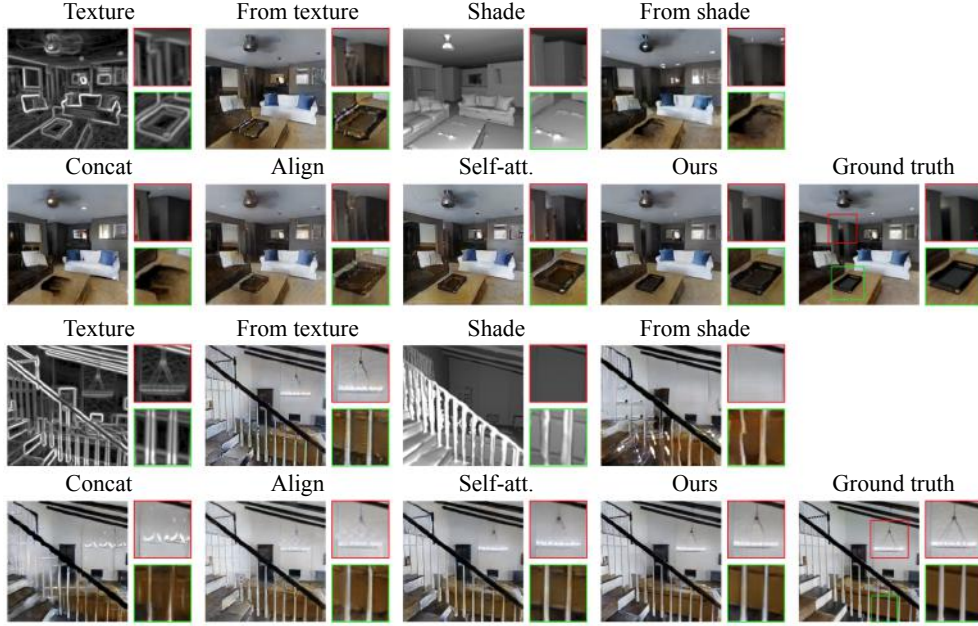


Figure 8: Two groups of results comparison for image translation from Texture and Shade to RGB. We observe that the prediction solely predicted from the texture is vague at boundary lines, while the prediction from the shade misses some opponents, *e.g.* the pendant lamp, and is weak in predicting handrails. When fusing the two modalities, the concatenation method is uncertain at the regions where both modalities have disagreements. Alignment and self-attention are still weak in combining both modalities at details. Our results are clear at boundaries and fine-grained details.

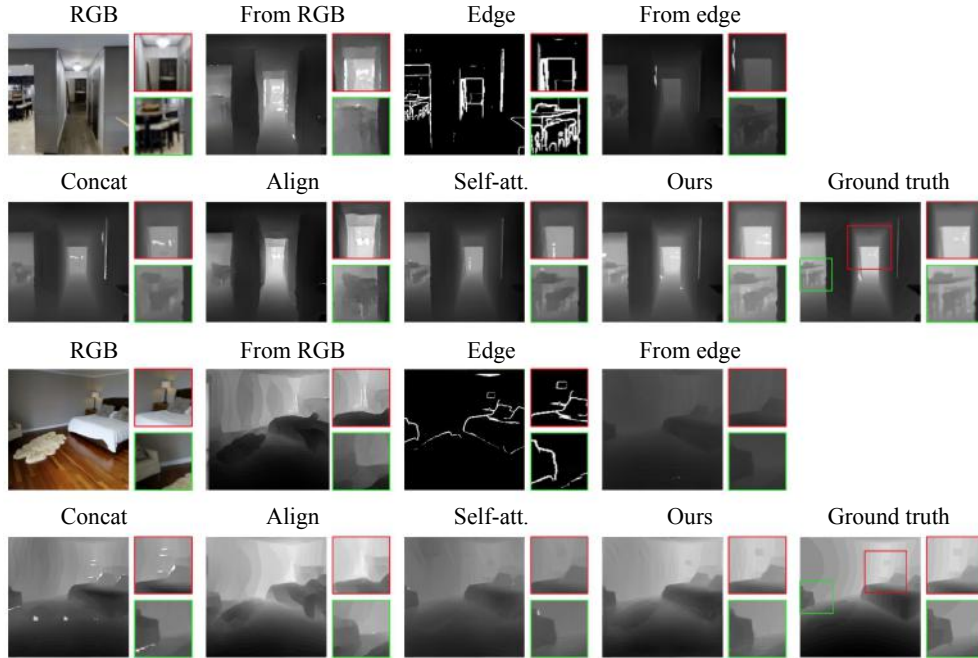


Figure 9: Two groups of results comparison for image translation from RGB and Edge to Depth. It is straightforward to find the benefits of multimodal fusion in this figure. The depth predicted by RGB is good at predicting numerical values, but is weak in capturing boundaries, which results in curving walls. Oppositely, the depth predicted by the edge well captures boundaries, but is weak in determining numerical values. The alignment fusion method is still weak in capturing boundaries. Both concatenation and self-attention methods are able to combine the advantages of both modalities, but the numerical values are still obviously lower than the ground truth. Our prediction achieves better performance compared to baseline methods.

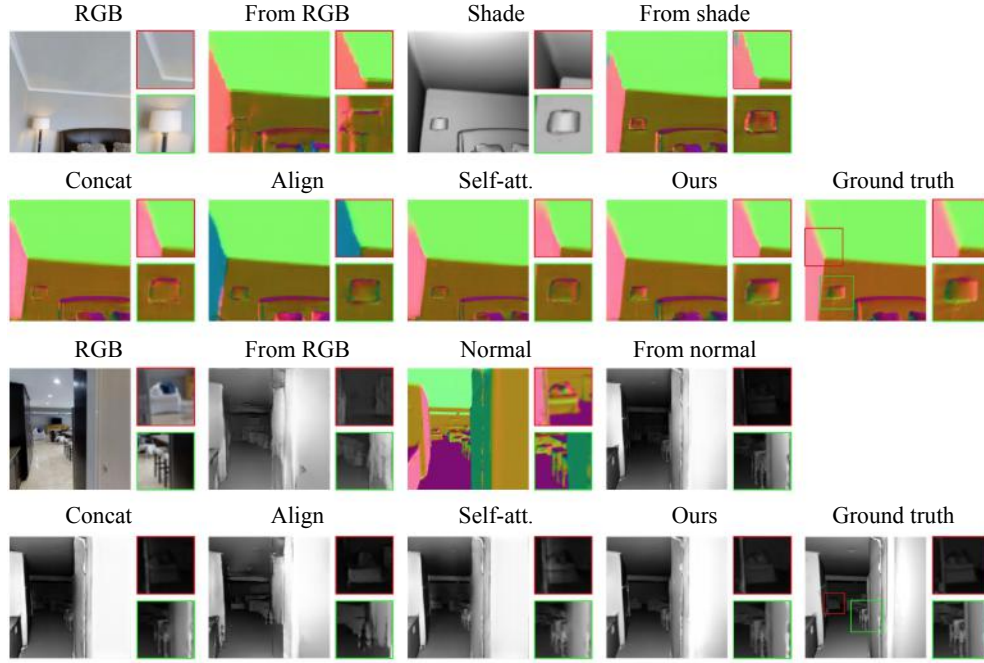


Figure 10: Results comparison for image translation from RGB and Shade to Normal (upper group), and from RGB and Normal to Shade (lower group). Our fusion method again outperforms the other methods regarding both overall performance and details.

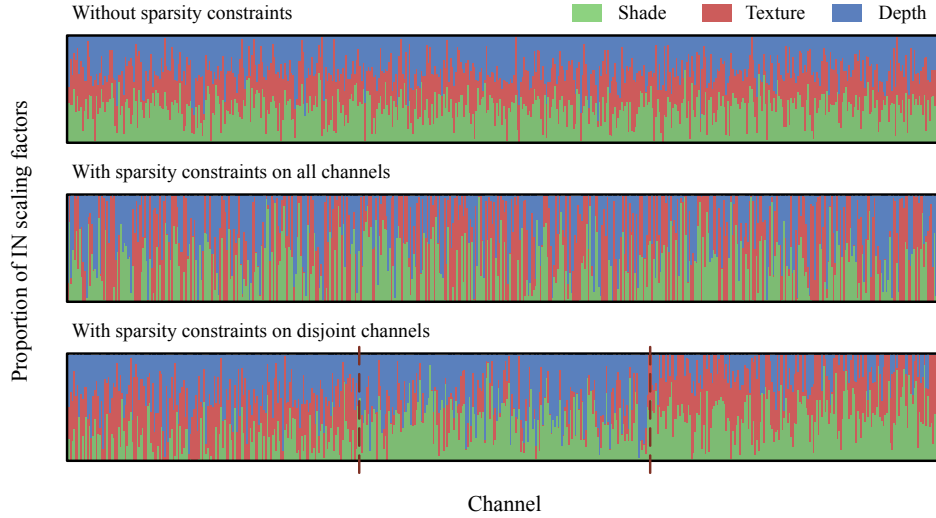


Figure 11: We use shared convs and unshared INs, and plot the proportion of scaling factors for each modality, at the 7th conv layer, *i.e.* $\gamma_c^{m,l,c} / (\gamma_c^{1,l,c} + \gamma_c^{2,l,c} + \gamma_c^{3,l,c})$, where $m = 1, 2, 3$ corresponding to Shade, Texture and Depth respectively, and $l = 7$. **Top:** no sparsity constraints are applied, where the scaling factor of each modality occupies a certain proportion at each channel. **Middle:** sparsity constraints are applied to all channels, where scaling factors of one modality could occupy a large proportion, indicating the channels are re-allocated to different modalities under the sparsity constraints. Yet this setting is not very suitable for channel exchanging, as a redundant feature map of one modality may be replaced by another redundant feature map. **Bottom:** sparsity constraints are applied on disjoint channels, which is our default setting.

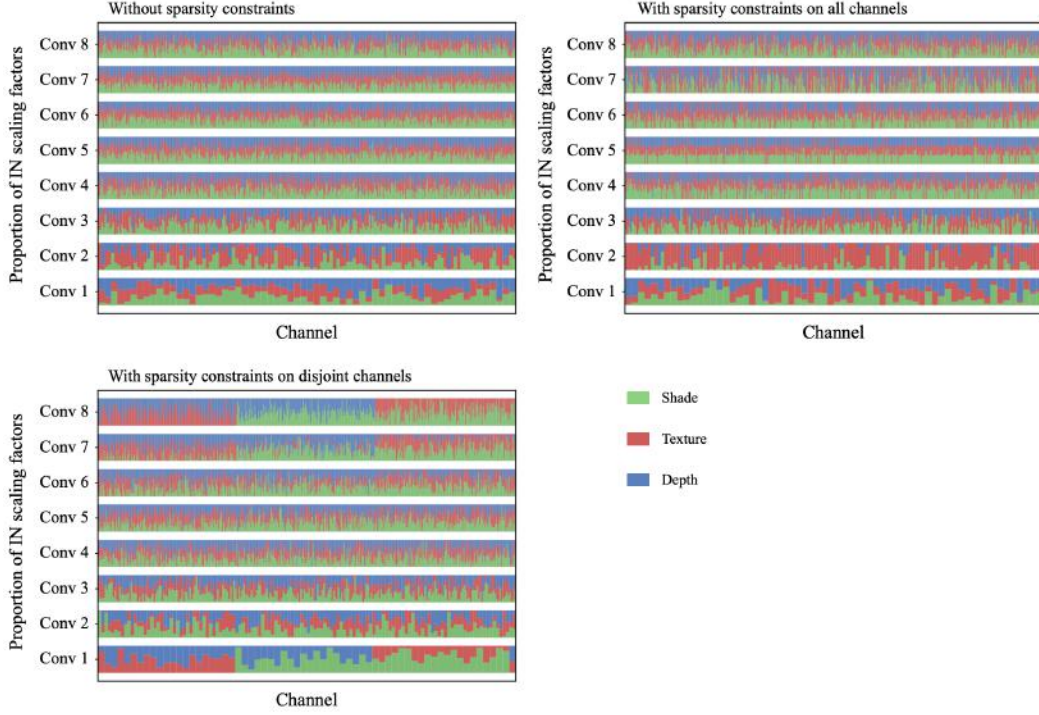


Figure 12: Proportion of scaling factors in the U-Net encoder. We provide results at all layers. **Upper left:** no sparsity constraints are applied; **Upper right:** sparsity constraints are applied on all channels; **Bottom left:** sparsity constraints are applied on disjoint channels.

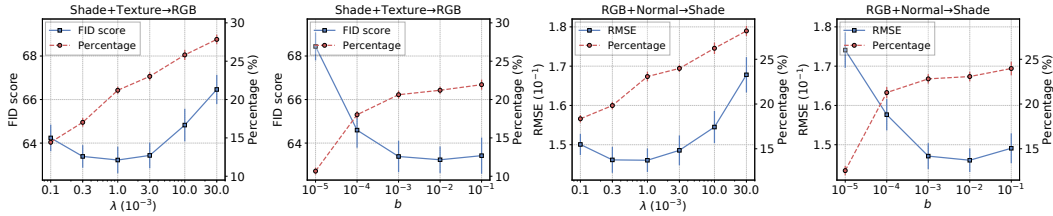


Figure 13: Sensitivity analysis for λ and θ . In our channel exchanging process, λ is the weight of sparsity constraint (Equation 4), and θ is the threshold for choosing close-to-zero scaling factors (Equation 6). We conduct five experiments for each parameter setting. In the 1st and 3rd sub-figures, λ ranges from 0.1×10^{-3} to 30.0×10^{-3} , and θ is set to 10^{-2} . In the 2nd and 4th sub-figures, θ ranges from 10^{-5} to 10^{-1} , and λ is set to 10^{-3} . The task name is shown at the top of each sub-figure. The left y-axis indicates the metric, and the right y-axis indicates the proportion of channels that are lower than the threshold θ , i.e. the proportion of channels that will be replaced. We observe that both hyper-parameters are not sensitive around their default settings ($\lambda = 10^{-3}$ and $\theta = 10^{-2}$).