FFNet: Frequency Fusion Network for Semantic Scene Completion

Xuzhi Wang[1], **Di Lin**[1], **Liang Wan**[1]*

[1]Colledge of Intelligence and Computing, Tianjin University
{wangxuzhi, lwan}@tju.edu.cn, ande.lin1988@gmail.com

Abstract

Semantic scene completion (SSC) aims at predicting the semantic categories and 3D volumetric occupancies of a scene simultaneously. The RGB-D images have been broadly used for providing the semantic and geometry information of the objects in existing semantic scene completion methods. However, existing works use concatenation, element-wise summation and weighted summation to fuse RGB-D data. These strategies ignore the large discrepancy of RGB-D data and the uncertainty measurements of depth data. To solve this problem, we propose the *Frequency Fusion Network* (FFNet), a novel method boosting semantic scene completion by better utilizing RGB-D data. It first explicitly correlates the RGB-D data in the frequency domain, which is different from the features directly extracted by the convolution operation. Then, the correlation information is used to guide the RGB-assisted depth features and depth-assisted RGB features. Further, considering the properties of different frequency components of RGB-D features, we propose a learnable elliptical mask to decompose the features, and attend to different frequency bands to facilitate the correlation process of RGB-D data. We evaluate FFNet intensively on the public SSC benchmarks, where FFNet surpasses the state-of-the-art methods. Our code will be made publicly available.

Introduction

Recent years have witnessed a great development of semantic scene completion for its applications in diverse fundamental tasks, e.g., grasping of robotics and obstacle avoidance of cars. Semantic scene completion aims to infer the 3D geometry occupancy of the scene as well as the semantic label of each voxel simultaneously (Song et al. 2017; Liu et al. 2018a; Zhang et al. 2019).

It has proven that RGB and depth data are useful cues in semantic scene completion task (Wang et al. 2019; Li et al. 2020b,d; Liu et al. 2018a). RGB images provide the semantic information which is vital for differentiating different objects. And depth images provide the geometry information which is important for inferring the 3D spatial structure and the 3D layout. Especially in some challenge indoor scenes, depth can boost semantic scene completion in poor lighting condition. RGB can help differentiate different objects with a close depth values.

Existing semantic scene completion methods usually adopt element-wise summation (Li et al. 2019, 2020b), element-wise multiplication (Li et al. 2020d), concatenation (Liu et al. 2018a) and weighted summation (Liu et al. 2020a) strategies to fuse the multi-modality data. However, these methods ignore the large discrepancy between the two modalities and there are many uncertainty measurements in the depth data, which is the challenge of RGB-D fusion (Chen et al. 2020c; Wang et al. 2020b; Valada, Mohan, and Burgard 2020; Piao et al. 2019). As a result, they are incapable of well utilizing the RGB and depth data to boost semantic scene completion task.

The pair of RGB and depth data are different representations of the same scene. The RGB data is the photometric representation and the depth data is the geometrical representation. They have a strong structural similarity (Fu et al. 2020; Chen et al. 2020c; Chen and Fu 2020). Inspired by this observation, we aim at better utilizing the correlation between RGB-D features for multi-modality fusion.

Motivated by the above analysis, we propose a novel and effective Frequency Fusion Network(FFNet) to realize RGB-D fusion for semantic scene completion. It adopts a correlation and assistance pipeline to tackle the challenges of RGB-D fusion. The key idea of the proposed method is to explicitly obtain the correlation of RGB-D using frequency learning and the correlation information is used to guide the generation of RGB-assisted depth features and depth-assisted RGB features. Our Frequency Fusion Module leads to the structure information enhancement for the RGB features and the depth features are recalibrated to fit for the RGB features.

To boost the correlation process of the RGB-D data, we further propose a novel mechanism that facilitates one modality better interacting with the other. We first decompose the RGB features and depth features by an elliptical mask which is learned and data-adaptive. The frequency describes the spatial changing of an image. High-frequency components correspond to the rapidly changing area such as the edge of an object. The low-frequency components represent the smoothly changing areas such as the body of an object (Li et al. 2020e). Intuitively, the similarity of different frequency components between RGB and depth features are

*Corresponding Author.

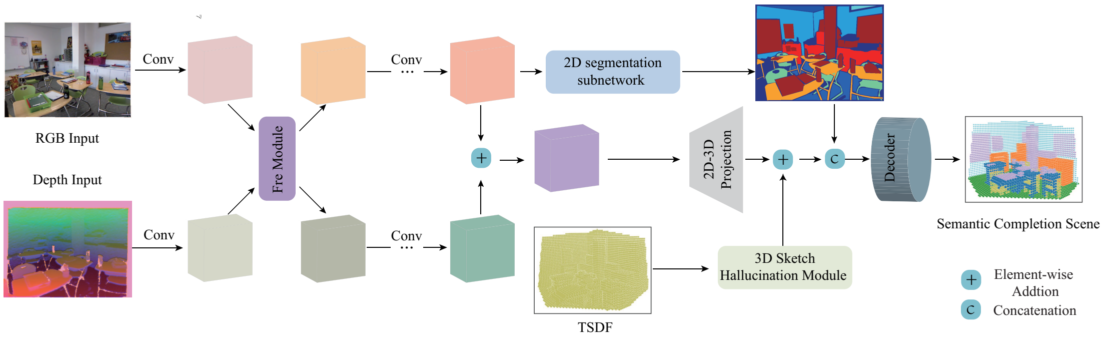

Figure 1: The overview framework of FFNet. The RGB and depth data are first fed into a ResNet-50 to extract features, and the RGB-D features are interacted and fused by the proposed Frequency Fusion Module. Then, the fused RGB-D features are mapped to 3D by the 2D-3D projection layer. The 2D segmentation subnetwork is used to predict the 2D semantic label from the features extracted by RGB branch. The semantic segmentation prediction is projected to 3D and concatenates with the projected RGB-D feature. In the end, the concatenated features are fed into the decoder to predict the semantic scene completion results. 'Fre Module' represents the proposed Frequency Fusion Module. The detail of the Frequency Fusion Module is illustrated in Figure 2.

different. Therefore, we aim at finding a more effective correlation between RGB-D features by emphasizing different frequency components.

There are two advantages of using frequency learning to correlate the RGB-D features. First, it can explicitly find the correlation of RGB-D data. Different from the features directly extracted by convolution operations, the proposed frequency correlation can be regarded as calculating the 'patch similarity' between one modality and every position of the other modality. It thus obtains the relation between RGB-D data. Second, emphasizing different frequency components decomposed by the learned elliptical mask can facilitate finding the correlation of RGB-D data.

Therefore, as shown in Figure 2 and 3, we carry out the sequential three steps to fuse RGB-D data: 1) emphasizing different frequency components of RGB-D features by elliptical decomposition mask, 2) correlating the RGB-D features in the frequency domain, 3) aggregating the correlation descriptor with RGB and depth features separately to generate the depth-assisted RGB features and RGB-assisted depth features.

As shown in Figure 1, the framework of FFNet comprises four stages: 1) a 2D feature extraction stage which extracts 2D multi-modality features boosted by Frequency Fusion Block, 2) a 2D segmentation subnetwork that uses the extracted RGB features to predict a 2D semantic segmentation mask and the 2D mask are concatenated with the fused RGB-D features, 3) a 2D-3D projection layer which projects the 2D feature to 3D volume and 4) 3D learning stage.

Our contribution can be summarized as follows:

- We propose a novel and effective method, termed Frequency Fusion Network, which can better utilize the complementary information of RGB-D data for semantic scene completion. It fuses the RGB-D features by explicitly model the correlation between the RGB and the depth data in the frequency domain. Then, the correlation descriptor is used to guide the generation of RGB-assisted depth feature and depth-assisted RGB feature.
- We propose a frequency attention mechanism to boost the correlation of the RGB-D feature. Frequency attention is achieved by attending to different frequencies of each modality using the proposed elliptical mask which is learnable and data-adaptive.
- The proposed method outperforms the state-of-the-art methods on the NYU and the NYUCAD dataset.

Related Work

Semantic Scene Completion Song et al. (Song et al. 2017) first focus on the task of semantic scene completion. They observe that occupancy patterns of the environment and the semantic labels of the objects are tightly intertwined. Therefore, they use 3D CNN to simultaneously predict the semantic labels and volumetric occupancy. Semantic scene completion methods can be mainly categorized into depth methods and RGB-D methods according to the input data.

For depth as input, most of the works encode the observed depth as Signed Distance Function(SDF). It can directly represent the 2D observation into the same 3D physical and facilitate the network to learn geometry and scene representation. ESSCNet (Zhang et al. 2018) speeds up semantic scene completion by Spatial Group Convolution which divides input voxels into different groups then carries out 3D sparse convolution on these separated groups. The work in (Zhang et al. 2019) proposes a cascaded context pyramid network and a guided residual refinement module to integrate both local geometric details and multi-scale 3D contexts of the scene.

Another category is using RGB-D as input (Li et al. 2020d; Liu et al. 2018a; Chen et al. 2020b; Li et al. 2020b; Cai et al. 2021). (Liu et al. 2018a) proposes a CNN-based framework that sequentially accomplishes two subtasks, i.e., 2D semantic segmentation and 3D semantic scene completion. They extract the RGB-D features in a double-branch

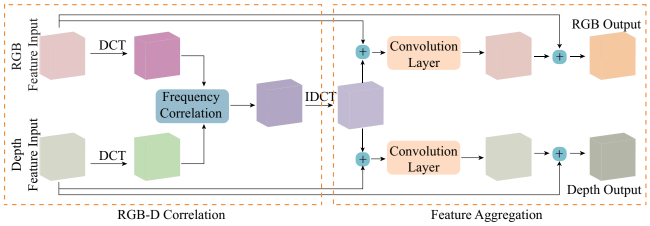

Figure 2: The overview of Frequency Fusion Module. It contains two main parts: RGB-D Correlation and Feature Aggregation. The details of Frequency Correlation is illustrated in Figure 3.

way and the extracted RGB-D features are fused by concatenation. (Chen et al. 2020b) present a novel anisotropic convolutional network that is much less computational demanding. Siqi et al. (Li et al. 2020d) propose the AMFNet that conducts 3D scene completion and semantic segmentation simultaneously via leveraging the experience of 2D segmentation and the reliable depth cues in the spatial dimension. The two tasks are fuse by element-wise multiplication. Yu et al. (Liu et al. 2020a) propose a 3D gated recurrent fusion network to fuse the information from depth and RGB.

Different from previous RGB-D SSC methods that combine RGB-D features by concatenation, element-wise summation and weighted summation. We propose the Frequency Fusion Network to explicitly model the correlation of RGB-D and then use the correlation information to fuse RGB-D data for semantic scene completion.

Frequency Domain Learning Frequency analysis has been widely used in signal processing and computer vision. Recently, many works inspired by frequency analysis and aim at endowing neural network with powerful ability by frequency domain learning (Jin et al. 2020; Helou, Zhou, and Susstrunk 2020; Qian et al. 2020; Liu et al. 2018b; Li et al. 2020c; Ehrlich and Davis 2019; Bian et al. 2020; Li et al. 2020e). Most of these works focus on using the compressing ability of Discrete Cosine Transform(DCT) which can decompose the input signal and discover their redundancy in the frequency domain. (Chen et al. 2020a) propose a frequency method for network pruning by converting filters into the frequency domain to investigate their redundancy. (Kai et al. 2020) propose to reshape the high-resolution images in the DCT domain and feed the reshaped DCT coefficients to neural networks. (Wang et al. 2020a) investigate the generalization ability of convolutional neural networks by frequency spectrum of image data.

Different from the above methods, our method obtains the correlation of the RGB-D data to guide RGB-D fusion in the frequency domain. And we propose the frequency attention by learned elliptical mask to facilitate modeling the correlation of RGB-D.

Method

In this section, we will present an overview of our architecture. Taking depth and its RGB counterpart as input, the network aims at predicting the 3D voxel occupancy and the semantic categories of each voxel as illustrated in Figure 1. Each voxel is mapped to one of the semantic labels $C = [c_0, c_1, ..., c_{N-1}]$, where C_0 resents the empty voxels and N is the number of semantic categories. Concretely, The RGB and depth data are first fed into a ResNet-50 to extract features, and the RGB-D features are interacted and fused by the proposed Frequency Fusion Module. Then, the fused RGB-D features are mapped to 3D by the 2D-3D projection layer. The 2D segmentation subnetwork is used to predict the 2D semantic label from the features extracted by RGB branch. The semantic segmentation prediction is projected to 3D and concatenates with the projected RGB-D feature. In the end, the concatenated features are fed into the decoder to predict the semantic scene completion results.

Next, we will introduce the details of the proposed methods from the following aspects: 1) Frequency Fusion Module, 2) RGB-D Semantic Scene Completion Framework, 3) Loss Function.

Frequency Fusion Module

Figure 2 shows the overall framework of RGB-D Frequency Fusion and Figure 3 shows the frequency correlation process of Frequency Fusion. The key idea of Frequency Fusion is to first obtain the explicit correlation of RGB-D feature and then the correlation descriptor is used to guide the generation of RGB-assisted depth feature and depth-assisted RGB features. We carry out the sequential three steps: 1) Frequency Attention, 2) RGB-D Frequency Correlation and 3) Feature Aggregation.

Frequency Attention The pipeline of frequency attention is shown in Figure 3. It aims at emphasizing different frequency components to boost RGB-D correlation by the mask of elliptical decomposition. Frequency is global information. The high-frequency components are related to the edges of an object and the low-frequency components are related to the body of an object. The RGB data and depth data are the photometric and geometrical representations of the same scene, respectively. The main idea of frequency attention is that we can facilitate the correlation between RGB-D by emphasizing different frequency components. Inspired that the 2D lowpass filter passes the frequencies within a circle of radius D in signal processing. We propose the learned elliptical mask to decompose the frequency signal for it is more flexible compared to a circle.

We denote $r_i \in \mathbb{R}^{H \times W}$ as the i-th channel of RGB features and denote $d_i \in \mathbb{R}^{H \times W}$ as the i-th channel of depth feature. We first transform the RGB and depth features into frequency domain by:

$$R_i = DCT(r_i), \tag{1}$$

$$D_i = DCT(d_i), \tag{2}$$

where R_i represents the RGB feature in the frequency domain, D_i represents the depth feature in the frequency domain and $DCT(\cdot)$ represents the discrete cosine transform that transforms a feature in the spatial domain to the frequency domain.

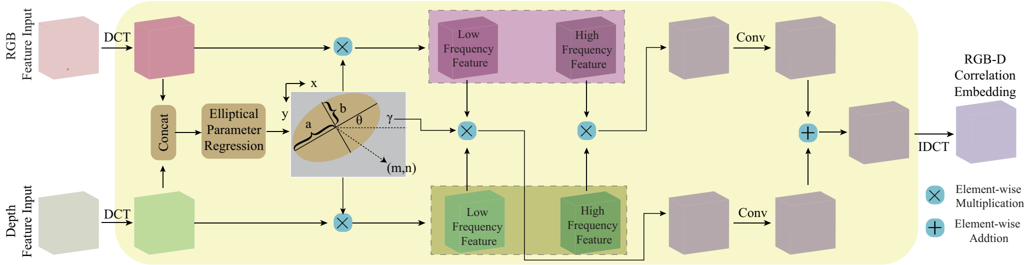

Figure 3: The overall architecture of frequency correlation boosted by frequency attention. Note that γ is the weight to emphasize the masked frequency component by elliptical decomposition.

Then, we use the elliptical mask to decompose each channel of RGB-D features in frequency domain into separate frequency components. The elliptical mask is learned by a parameter prediction network which regresses the five parameters of the elliptical mask and a parameter of frequency attention. The frequency attention parameter is used to emphasize different frequency components. The detail of the parameter prediction network can be found in the supplementary material.

The elliptical mask of i-th channel is defined as:

$$M_i = \begin{cases} 1, \frac{((x-m)cos\theta-(y-m)sin\theta)^2}{a^2} - \frac{((x-n)cos\theta-(y-n)sin\theta)^2}{b^2} < 1 \\ 0, otherwise \end{cases} \tag{3}$$

where M_i represents the elliptical mask of the i-th channel, (m, n) represents the elliptical center in the elliptical mask, a represents the semi-major axis, b represents the semi-minor axis and θ represents the rotation angle.

Then, we use the mask weight γ to emphasize or suppress the masked frequency component. γ is learned by the parameter prediction network. The output of frequency attention can be written as:

$$R^{'} = Conv(\sigma(\gamma \cdot M \cdot R)) + Conv(\sigma(\overline{M} \cdot R)), \tag{4}$$

$$D^{'} = Conv(\sigma(\gamma \cdot M \cdot D)) + Conv(\sigma(\overline{M} \cdot D)), \tag{5}$$

where R represents the frequecy information transfromed from RGB features by DCT, D represents the frequency information transformed from depth features by DCT, $R^{'}$ represents the frequency enhanced RGB features in DCT domain, $D^{'}$ represents the frequency enhanced depth features in DCT domain, $\overline{M}$ = $1 - M$, $M = [M_1, M_2, \cdots, M_C]$ and γ represents the weight of low frequency component.

RGB-D Frequency Correlation RGB-D correlation aims at explicitly finding the correlation of RGB-D features. To correlate RGB features with depth features, we define an operation in the frequency domain:

$$I = R^{'} \cdot D^{'}, \tag{6}$$

where $\cdot$ represents pixel-wise multiplication and I is the RGB-D correlation information.

Then, the correlation information of RGB-D features is normalized and learned by:

$$I^{'} = Conv(\sigma(I)), \tag{7}$$

where $\sigma(\cdot)$ represents element-wise sigmoid fuction and $Conv(\cdot)$ represents the convolution layer, $I^{'}$ is the normalized and learned features of I and $I = [I_1, I_2, \cdots, I_C]$. C represents the channel dimension.

Then, the $I^{'}$ is transformed into spatial domain by:

$$F_i^{cor} = IDCT(I_i^{'}), \tag{8}$$

where $F_i^{cor} \in \mathbb{R}^{H \times W}$ represents the RGB-D correlation embedding in the spatial domain, $IDCT(\cdot)$ represents inverse discrete cosine transform that transforms a feature in the frequency domain to the spatial domain.

We further analyze the RGB-D frequency correlation by:

$$r_i * d_i = IDCT(DCT(r_i) \cdot DCT(d_i)), \tag{9}$$

Where $*$ represents convolution operation, r_i represents the i-th channel of RGB feature and d_i represents the i-th channel of depth feature. Notably, it is the well-known convolution theorem that point-wise multiplication in the frequency domain equals convolution in the spatial domain. So the above operation in frequency domain equals to convolve depth feature with RGB features in the spatial domain, in other words, choose one of them as a convolution kernel and convolve with the other. Moreover, We add convolution operation to adjust and learn the features to facilitate RGB-D correlation. It thus obtains an explicit correlation of RGB-D features.

Feature Aggregation The feature aggregation process is shown in Figure 2. To make full use of the complementarity of RGB-D features, we take the correlation information of RGB-D features as guidance to generate the RGB-assisted depth feature and depth-assisted RGB feature separately.

To make full use of the complementarity of the RGB-D feature, we define this operation:

$$r_{att} = Conv(F^{cor} + r), \tag{10}$$

$$d_{att} = Conv(F^{cor} + d), \tag{11}$$

where r_{att} represents the attention score generated by RGB features and the correlation information of RGB-D, d_{att} represents the attention score generated by depth features and the correlation information of RGB-D and $Conv$ represents the convolution operation.

The final output of Frequency Fusion Module can be written as:

$$r_{out} = r + r_{att} \tag{12}$$

$$d_{out} = d + d_{att} \tag{13}$$

where r_{out} represents the depth-assisted RGB features, d_{out} represents the RGB-assisted depth features and $+$ denotes a residual connection, which allows us to insert our block into any network, without breaking its initial behavior.

Semantic Scene Completion Framework

We take the RGB-D images as input and predict the semantic labels of the 3D scenes. The framework of semantic scene completion contains four stages: 2D feature extraction, 2D semantic segmentation, 2D-3D projection and 3D feature learning.

2D feature extraction To extract 2D features from the depth and RGB image, we use a ResNet-50 model which is pre-training on ImageNet to the RGB branch. The weight of the RGB branch is fixed and the weight of the depth branch is not fixed. Fixing the weights can better utilize the pre-trained ResNet50 RGB features for SSC. Chen et al. (Chen et al. 2020b) also use this setting. And for RGB-D fusion, we fix the weights of RGB branch for stable training. The two branches are interacting and fusing by the proposed Frequency Fusion Module.

2D Semantic Segmentation The aim of 2D semantic segmentation is to acquiring pixel-wise semantic predictions to boost the semantic scene completion task. We use the RGB branch of the 2D feature extraction subnetwork as our encoder which is a pre-trained ResNet-50. We use the decoder as DeepLab v3+ (Chen et al. 2018). DeepLab v3+ is first pre-trained on the ADE-20k dataset (Zhou et al. 2017) and finetuned on the NYU dataset. The segmentation results are projected to the corresponding 3D space using the 2D-3D projection layer.

2D-3D projection To alleviate the gap between 2D and 3D, the fused 2D RGB-D features and 2D semantic segmentation results are projected into the corresponding 3D positions by 2D-3D projection layer according to the intrinsic camera matrix K_{camera}, the extrinsic camera matrix $[R|t]$ and the depth image I_{depth}.

3D feature learning We add the 3D features projected from the 2D fused RGB-D features and 3D sketch predicted by 3D Sketch Hallucination Module (Chen et al. 2020b). The 3D sketch could be understood as a kind of 3D boundary, which could encode the geometric information. Then the added features are concatenated with the projected 2D semantic segmentation predictions. In the following, the concatenated 3D features are fed into a stacked of AIC modules (Li et al. 2019), which is a lightweight 3D CNN module. Finally, we obtain the semantic scene completion results.

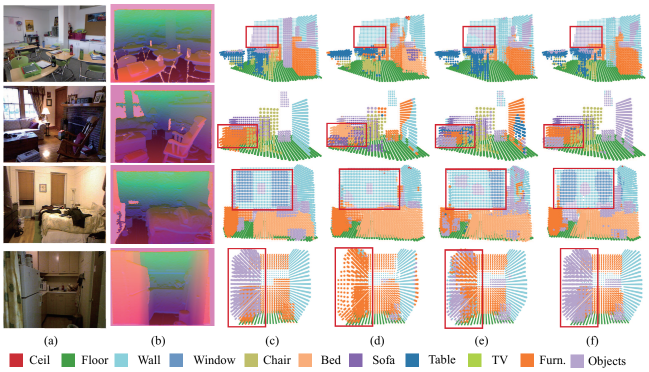

Figure 4: SSC results with different methods on the NYU dataset. From left to right: (a) Input RGB; (b) Input Depth (HHA); (c) Ground truth; (d) SSCNet; (e) Our method without Frequency Fusion; (f) Our method with Frequency Fusion.

Loss Function

Given RGB-D images and ground truth semantic labels of the 3D scenes, our proposed method can be trained in an end-to-end manner. We jointly supervise the two parts, including L_{sm} and L_{sk} (Chen et al. 2020b). The total loss L is computed as:

$$L = L_{sm} + L_{sk}, \tag{14}$$

where L_{sm} represents the semantic loss, and L_{sk} (Chen et al. 2020b) represents the sketch loss. We adopt the voxel-wise cross-entropy loss function for the network training. The semantic loss function can be written as:

$$L_{sm} = \sum_{ijk} \omega_{ijk} L_{sm}(\hat{y}_{ijk}, y_{ijk}), \tag{15}$$

where $\hat{y}_{ijk}$ represents the predicted probability for the indexed voxel, y_{ijk} is the ground truth label, and ω_{ijk} represents the weight of each semantic category. We follow (Chen et al. 2020b) to use the sketch loss to supervise the sketch of each scene. A sketch refers to the 3D boundary of a 3D scene.

Experiments

Implementation Details

Given the training data (i.e. the RGB image, the depth image and the ground truth 3D labels), we train our network in an end-to-end manner. We implement our framework in PyTorch. We train our model with batch size 6 in 2 GeForce GTX 3090 Ti GPUs. We adopt mini-batch S-GD with momentum 0.9 and weight decay 0.0005. For both NYU and NYU CAD datasets, we train our network for 350 epochs with initial learning rate 0.1. We use a poly learning rate policy where the initial learning rate is updated by $(1 - \frac{iteration}{max\_iteration})^{0.9}$. The inference time is 0.58 seconds per scene.

The introduction of the dataset and evaluation metrics are detailed in the supplementary material.

		scene completion			semantic scene completion											
Methods	Trained on	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SSCNet (Song et al. 2017)	NYU	57.0	**94.5**	55.1	15.1	94.7	24.4	0.0	12.6	32.1	35.0	13.0	7.8	27.1	10.1	24.7
ESSCNet (Zhang et al. 2018)	NYU	71.9	71.9	56.2	17.5	75.4	25.8	6.7	15.3	53.8	42.4	11.2	0	33.4	11.8	26.7
DDRNet (Li et al. 2019)	NYU	71.5	80.8	61.0	21.1	92.2	33.5	6.8	14.8	48.3	42.3	13.2	13.9	35.3	13.2	30.4
AICNet (Li et al. 2020b)	NYU	62.4	91.8	59.2	23.2	90.8	32.3	14.8	18.2	51.1	44.8	15.2	22.4	38.3	15.7	33.3
TS3D (Liu et al. 2018a)	NYU	-	-	60.0	9.7	93.4	25.5	21.0	17.4	55.9	49.2	17.0	27.5	39.4	19.3	34.1
PALNet (Liu et al. 2020b)	NYU	68.7	85.0	61.3	23.5	92.0	33.0	11.6	20.1	53.9	48.1	16.2	24.2	37.8	14.7	34.1
AMFNet (Li et al. 2020d)	NYU	66.3	80.5	57.2	20.0	78.7	27.3	20.5	21.8	56.5	53.9	19.5	18.8	40.1	19.5	34.2
CCPNet (Zhang et al. 2019)	NYU	74.2	90.8	63.5	23.5	96.3	35.7	20.2	25.8	61.4	56.1	18.1	28.1	37.8	20.1	38.5
Chen et al. (Chen et al. 2020b)	NYU	85.0	81.6	71.3	43.1	93.6	40.5	24.3	30.0	57.1	49.3	**29.2**	14.3	42.4	28.6	41.1
CCPNet (Li et al. 2019)	NYU+SUNCG	78.8	94.3	67.1	25.5	**98.5**	38.8	27.1	27.3	**64.8**	**58.4**	21.5	**30.1**	38.4	23.8	41.3
Ours	NYU	**89.3**	78.5	**71.8**	**44.0**	93.7	**41.5**	**29.3**	**36.2**	59.0	51.1	28.9	26.5	**45.0**	**32.6**	**44.4**

Table 1: Results on NYU dataset.

		scene completion			semantic scene completion											
Methods	Trained on	prec.	recall	IoU	ceil.	floor	wall	win.	chair	bed	sofa	table	tvs	furn.	objs.	avg.
SSCNet (Song et al. 2017)	NYUCAD+SUNCG	75.4	96.3	73.2	32.5	92.6	40.2	8.9	33.9	57.0	59.5	28.3	8.1	44.8	25.1	40.0
DDRNet (Li et al. 2019)	NYUCAD	88.7	88.5	79.4	54.1	91.5	56.4	14.9	37.0	55.7	51.0	28.8	9.2	44.1	27.8	42.8
AICNet (Li et al. 2019)	NYUCAD	88.2	90.3	80.5	53.0	91.2	57.2	20.2	44.6	58.4	56.2	36.2	9.7	47.1	30.4	45.8
TS3D (Liu et al. 2018a)	NYUCAD	-	-	76.1	25.9	93.8	48.9	33.4	31.2	66.1	56.4	31.6	**38.5**	51.4	30.8	46.2
PALNet (Liu et al. 2020b)	NYUCAD	87.2	91.7	80.8	54.8	92.8	60.3	15.3	43.1	60.7	59.9	37.6	8.1	48.6	31.7	46.6
AMFNet (Li et al. 2020d)	NYUCAD	60.6	89.1	56.3	81.3	68.5	54.1	61.8	30.2	45.9	50.7	34.3	42.7	41.9	28.4	49.1
CCPNet (Zhang et al. 2019)	NYUCAD	91.3	92.6	82.4	56.2	94.6	58.7	35.1	44.8	68.6	65.3	37.6	35.5	53.1	35.2	53.2
Chen et al. (Chen et al. 2020b)	NYUCAD	90.6	92.2	84.2	59.7	94.3	64.3	32.6	51.7	72.0	68.7	45.9	19.0	60.5	**38.5**	55.2
CCPNet (Zhang et al. 2019)	NYUCAD+SUNCG	93.4	91.2	85.1	58.1	**95.1**	60.5	**36.8**	47.2	69.3	67.7	39.8	37.6	55.4	37.6	55.5
Ours	NYUCAD	**94.8**	90.3	**85.5**	**62.7**	**94.9**	**67.9**	**35.2**	**52.0**	**74.8**	**69.9**	**47.9**	27.9	**62.7**	35.1	**57.4**

Table 2: Results on NYU CAD dataset

Method	mIoU
ResNet-50	42.1
ResNet-50+ cmFM +AFS (Li et al. 2020a)	42.8
ResNet-50+ SA-Gate (Chen et al. 2020c)	43.2
ResNet-50+Fre-Fusion	**44.4**

Table 3: The proposed Frequency Fusion Module vs. other RGB-D fusion module

Comparisons with the State-of-the-art Methods

Quantitative Comparision We compare the proposed method on NYU and NYU CAD dataset with state-of-the-art methods. Table 1 shows the results on NYU dataset. Our method achieves the best performance in both SC and SSC tasks. In spite of only taking NYU dataset for training, our approach obtains higher IoUs than CCP. Note that CCP uses supplementing training data from SUNCG dataset. We obtain an improvement of 4.7% SC IoU and 3.1% SSC mIoU compared to CCP method. This indicates that our method can better exploit the complementary information from the RGB and depth data for semantic scene completion.

Table 2 shows the results on NYU CAD dataset. Our method achieves the best performance in both SC and SSC tasks. We obtain a improvement of 1.3% SC IoU and 2.2% SSC mIoU compared with (Chen et al. 2020b). Chen et al. only use NYU CAD as training data which is the same as us.

Qualitative Comparision Fig 4 shows the qualitative results on NYU dataset. We show the visualization results of our method, our method without Frequency Fusion Module and SSCNet. We can observe that the semantic scene completion result is more accurate with the proposed Frequency Fusion Module. Boosted by Frequency Fusion Module, our method shows an improvement in intra-class consistency and a sharper boundary between inter-class. Especially in the case that different objects have a similar depth, RGB data can provide complementary semantic features to differentiate different objects. It can be seen in the first, third and fourth rows that different objects have similar depth, RGB data can provide complementary semantic features to differentiate different objects. We can observe in the second row that the table in the bottom left corner has poor lighting conditions. It is difficult for the network to differentiate the table using only RGB images. By better utilizing the RGB-D data, our method can obtain a better semantic scene completion result.

Ablation Study

In this section, we evaluate the effectiveness of our core component, Frequency Fusion Module, in detail.

The Effectiveness of Using Frequency Fusion Module We insert Frequency Fusion Module into our ResNet-50 backbone to fuse the RGB and depth data. Here, we insert the Frequency Fusion Module before the first bottleneck of

Concat	Add	Prod	Channel	N-L	Proposed	mIoU
✓						41.4
	✓					43.2
		✓				43.3
			✓			43.5
				✓		43.7
					✓	**44.4**

Table 4: Ablation experiments on RGB-D Frequency Correlation part of Frequency Fusion Module on NYU dataset.

Fre-Corr	Fre-Atten	Aggregation	mIoU
			42.1
✓			43.2
✓	✓		43.8
✓	✓	✓	**44.4**

Table 5: Ablation experiments on Frequency Correlation, Frequency Attention and Feature Aggregation.

ResNet-50. To verify the effectiveness of our Frequency Fusion Module, first, we conduct the baseline method that does not use Frequency Fusion Module. Then, we replace the Frequency Fusion Module with other state-of-the-art RGB-D fusion methods used in RGB-D segmentation (Chen et al. 2020c) and Saliency detection (Li et al. 2020a). The results shown in Table 3 indicate that our method is effective and outperform other state-of-the-art RGB-D fusion methods.

The Effectiveness of Using Frequency Correlation Boosted by Frequency Attention We use Frequency Correlation to correlate RGB-D features boosted by frequency attention. To verify the effectiveness of this operation, we replace Frequency Correlation with five different architectures for comparison. ’Concat’ represents that we concatenate the RGB and depth features and use convolution to transform the features into the channel size the same as a single modality. ’Add’ means that we add RGB and depth feature to replace the frequency correlation information. ’Product’ represents that we multiply RGB features by depth features. ’Channel’ means that we use channel-wise attention (Hu, Shen, and Sun 2017) to fuse RGB-D features. ’N-L’ means that we use the non-local operation (Chi et al. 2020; Wang et al. 2018) to replace the frequency correlation information. Note that following the listed five architectures we use Feature Aggregation to obtain depth-assisted RGB features and RGB-assisted depth features. Table 4 shows that the proposed Frequency Correlation outperforms other architecture. Besides, the addition, production, channel-wise attention and non-local architecture can only promote a little performance. Concatenation will lead to relatively worse performance.

The Effectiveness of the three core components of Frequency Fusion Module We evaluate the effectiveness of the three core components of Frequency Fusion Module. We ablate each design of Frequency Fusion Module in Table 5. ’Fre-Atten’ represents that we use Frequency Attention to boost Frequency Correlation. ’Fre-Corr’ represents that we correlate the RGB-D features in the frequency domain. ’Aggregation’ represents that we aggregate the correlation information with RGB and depth features separately. Experiment results in Table 5 show the effectiveness of Frequency Attention, RGB-D Frequency Correlation and Feature Aggregation operations.

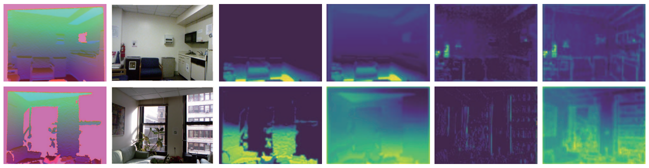
(a) HHA input (b) RGB input (c) HHA$_{in}$ (d) HHA$_{out}$ (e) RGB$_{in}$ (f) RGB$_{out}$

Figure 5: Visualization of depth and RGB feature before and after the Frequency Fusion Module. From left to right: (a) Depth(HHA) input; (b) RGB input; (c) HHA input to Frequency Fusion Module; (d) HHA output to Frequency Fusion Module; (e) RGB input to Frequency Fusion Module; (f) RGB output to Frequency Fusion Module.

How does Frequency Fusion help?

We show several representative feature samples before and after the proposed Frequency Fusion Module, to show how our method helps RGB-D fusion and semantic scene completion.

We first analyze the RGB features on the NYU dataset. As shown in Figure 5, it can be seen in the fifth and sixth column that the structure information is enhanced after the proposed module and thus reduce the background distraction. Structure information contains more geometry information which is vital for semantic scene completion.

Then, we analyze the HHA features on the NYU dataset. As shown in Figure 5, it can be seen that in the third and fourth columns that the proposed module can recalibrate the depth features to fit for the RGB feature.

In conclusion, the proposed Frequency Fusion Network could help find the correlation of RGB-D features and make full use of the advantages of RGB-D features to supplement each other.

Conclusion

RGB and depth data can provide complementary information which is an important cue for semantic scene completion. To better utilize the RGB-D data for semantic scene completion, we propose a novel and effective method, the Frequency Fusion Network. The proposed method can explicitly model the correlation of RGB-D features and the correlation is used to guide the RGB-assisted depth features and the depth-assisted RGB features. And we boost the correlation of RGB-D by frequency attention. It can thus alleviate the challenge of RGB-D fusion. Experiment results show that our method outperforms the state-of-the-art methods on NYU and NYU CAD dataset. Ablation study and visualization results also show the contribution of the proposed method.

Acknowledgments

This work was supported by the Natural Science Foundation of China (NSFC) under Grant 62072334.

References

Bian, S.; Wang, T.; Hiromoto, M.; and Shi, Y. 2020. ENSEI: Efficient Secure Inference via Frequency-Domain Homomorphic Convolution for Privacy-Preserving Visual Recognition. In *CVPR*.

Cai, Y.; Chen, X.; Zhang, C.; Lin, K.-Y.; Wang, X.; and Li, H. 2021. Semantic Scene Completion via Integrating Instances and Scene in-the-Loop. In *CVPR*.

Chen, H.; Wang, Y.; Shu, H.; Tang, Y.; and Xu, C. 2020a. Frequency domain compact 3D convolutional neural networks. In *CVPR*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *ECCV*.

Chen, S.; and Fu, Y. 2020. Progressively Guided Alternate Refinement Network for RGB-D Salient Object Detection. In *ECCV*.

Chen, X.; Lin, K.-Y.; Qian, C.; Zeng, G.; and Li, H. 2020b. 3D Sketch-aware Semantic Scene Completion via Semi-supervised Structure Prior. In *CVPR*.

Chen, X.; Lin, K.-Y.; Wang, J.; Wu, W.; Qian, C.; Li, H.; and Zeng, G. 2020c. Bi-directional Cross-Modality Feature Propagation with Separation-and-Aggregation Gate for RGB-D Semantic Segmentation. In *ECCV*.

Chi, L.; Yuan, Z.; Mu, Y.; and Wang, C. 2020. Non-Local Neural Networks with Grouped Bilinear Attentional Transforms. In *CVPR*.

Ehrlich, M.; and Davis, L. 2019. Deep Residual Learning in the JPEG Transform Domain. In *ICCV*.

Fu, K.; Fan, D.; Ji, G.; and Zhao, Q. 2020. JL-DCF:Joint Learning and Densely-Cooperative Fusion Framework for RGB-D Salient Object Detection. In *CVPR*.

Helou, M. E.; Zhou, R.; and Susstrunk, S. 2020. Stochastic Frequency Masking to Improve Super-Resolution and Denoising Networks. In *ECCV*.

Hu, J.; Shen, L.; and Sun, G. 2017. Squeeze-and-Excitation Networks. In *CVPR*.

Jin, B.; Hu, Y.; Tang, Q.; Niu, J.; Shi, Z.; Han, Y.; and Li, X. 2020. Exploring Spatial-Temporal Multi-Frequency Analysis for High-Fidelity and Temporal-Consistency Video Predition. In *CVPR*.

Kai, X.; Minghai, Q.; Fei, S.; and Yuhao, W. 2020. Learning in the frequency domain. In *CVPR*.

Li, C.; Cong, R.; Piao, Y.; Xu, Q.; and Loy, C. C. 2020a. RGB-D Salient Object Detection with Cross-Modality Modulation and Selection. In *ECCV*.

Li, J.; Han, K.; Wang, P.; Liu, Y.; and Yuan, X. 2020b. Anisotropic Convolutional Networks for 3D Semantic Scene Completion. In *CVPR*.

Li, J.; Liu, Y.; Gong, D.; Shi, Q.; Yuan, X.; and Zhao, C. 2019. RGBD Based Dimensional Decomposition Residual Network for 3D Semantic Scene Completion. In *CVPR*.

Li, S.; Xue, K.; Zhu, B.; Ding, C.; Gao, X.; Wei, D.; and Wan, T. 2020c. FALCON: A Fourier Transform Based Approach for Fast and Secure Convolutional Neural Network Predictions. In *CVPR*.

Li, S.; Zou, C.; Li, Y.; Zhao, X.; and Gao, Y. 2020d. Attention-based Multi-modal Fusion Network for Semantic Scene Completion. In *AAAI*.

Li, X.; Li, X.; Zhang, L.; Cheng, G.; Shi, J.; Liu, Z.; Tan, S.; and Tong, Y. 2020e. Improving Semantic Segmentation via Decoupled Body and Edge Supervision. In *CVPR*.

Liu, S.; Hu, Y.; Zeng, Y.; Tang, Q.; Jin, B.; Han, Y.; and Li, X. 2018a. See and think: Disentangling semantic scene completion. In *NIPS*.

Liu, Y.; Li, J.; ; Yan, Q.; Yuan, X.; Zhao, C.; Reid, I.; and Cadena, C. 2020a. 3D Gated Recurrent Fusion for Semantic Scene Completion. In *arXiv: 2002.07269*.

Liu, Y.; Li, J.; Yuan, X.; Zhao, C.; Siegwart, R.; Reid, I.; and Cadena, C. 2020b. Depth Based Semantic Scene Completion with Position Importance Aware Loss. In *ICRA*.

Liu, Z.; Xu, J.; Peng, X.; and Xiong, R. 2018b. Frequency-Domain Dynamic pruning for Convolutional Neural Networks. In *NIPS*.

Piao, Y.; Ji, W.; Li, J.; Zhang, M.; and Lu, H. 2019. Depth-induced Multi-scale Recurrent Attention Network for Saliency Detection. In *ICCV*.

Qian, Y.; Yin, G.; Sheng, L.; Chen, Z.; and Shao, J. 2020. Thinking in Frequency: Face Forgery Detection by Mining Frequency-aware Clues. In *ECCV*.

Song, S.; Yu, F.; Zeng, A.; Chang, A. X.; Savva, M.; and Funkhouser, T. 2017. Semantic scene completion from a single depth image. In *CVPR*.

Valada, A.; Mohan, R.; and Burgard, W. 2020. Self-Supervised Model Adaptation for Multimodal Semantic Segmentation. *International Journal of Computer Vision*, 128(3).

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020a. High frequency component helps explain the generalization of convolutional neural networks. In *CVPR*.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local Neural Networks. In *CVPR*.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020b. ABMDRNet: Adaptive-weighted Bi-directional Modality Difference Reduction Network for RGB-T Semantic Segmentation. In *NIPS*.

Wang, Y.; Tan, D. J.; Navab, N.; and Tombari, F. 2019. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *ICCV*.

Zhang, J.; Zhao, H.; Yao, A.; Chen, Y.; Zhang, L.; and Liao, H. 2018. Efficient semantic scene completion network with spatial group convolution. In *ECCV*.

Zhang, P.; Liu, W.; Lei, Y.; Lu, H.; and Yang, X. 2019. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *ICCV*.

Zhou, B.; Zhao, H.; Puig, X.; Fidler, S.; Barriuso, A.; and Torralba, A. 2017. Scene Parsing through ADE20k Dataset. In *CVPR*.