IET Journals
The Institution of Engineering and Technology

FAFNet: Fully-Aligned Fusion Network for RGBD Semantic Segmentation based on Hierarchical Semantic Flows

ISSN 1751-8644
doi: 0000000000
www.ietdl.org

*Jiazhou Chen, Yangfan Zhan, Yanghui Xu, Xiang Pan**

School of Computer Science and Technology, Zhejiang University of Technology, 288 Liuhe Road, Hangzhou, P.R.China
** E-mail: panx@zjut.edu.cn*

Abstract: Depth maps are acquirable and irreplaceable geometric information that significantly enhances traditional color images. RGB and Depth (RGBD) images have been widely used in various image analysis applications, but they are still very limited due to challenges from different modalities and misalignment between color and depth. In this paper, we present a Fully-Aligned Fusion Network (FAFNet) for RGBD semantic segmentation. To improve cross-modality fusion, a new RGBD fusion block is proposed, features from color images and depth maps are first fused by an attention cross fusion module and then aligned by a semantic flow. A multi-layer structure is also designed to hierarchically utilize our RGBD fusion block, which not only eases issues of low resolution and noises for depth maps but also reduces the loss of semantic features in the upsampling process. Quantitative and qualitative evaluations on both the NYU-Depth V2 and the SUN RGB-D dataset demonstrate that our FAFNet model outperforms state-of-the-art RGBD semantic segmentation methods.

1 Introduction

With the rapid development of deep learning techniques, semantic segmentation has become an important research direction in the field of computer vision. It refers to the process of segmenting an image into regions that belong to different classes at a pixel level. Beyond the region segmentation, semantic segmentation predicts the classification for each region through deep learning techniques[1–6], thus it can provide a higher-level understanding of objects in the input image than traditional image segmentation methods that are based on low-level features[7, 8].

Although learning-based semantic segmentation has made great progress in the last decade, it is still limited due to the lack of geometric information in the color image. For instance, it can neither separate different objects sharing the same boundary and high color similarities nor recognize objects that contains very complex textures.

A depth map is an image containing information related to the distance to the object's surface from the point of view. Though depth is not full 3D geometry, it provides abundant geometric information of objects that is hardly presented in color images, thus it is widely used in various applications, such as shape completion[9], drone navigation[10], etc. With the popularization of commercial depth sensors, such as Microsoft Kinect and Intel RealSence, depth maps have become cheaper and easier to acquire and can upgrade RGB color images to RGBD images. Therefore, more and more neural networks in the literature have been proposed for RGBD semantic segmentation[11–15]. Though RGBD-based methods significantly improve the segmentation accuracy, fusing depth maps with color images in the semantic segmentation task is still very challenging due to three main reasons:

1) Color and depth are completely different modalities, directly combining color pixels and depth ones by channel concatenating (to RGBD pixels) may introduce some ambiguities, as the neural network can not establish a sufficient correlation between two irrelevant modalities.

2) Color and depth pixels are not fully aligned, since they are captured by different lens[16]. Though image registration is usually taken, the misalignment is still unavoidable, especially for consumer RGBD cameras, such as Microsoft Kinect.

3) Finally, depth maps always have lower resolution and more noises than color images. Depth maps may mislead the semantic segmentation if low-quality depth features are fused to color features.

To overcome the aforementioned challenges, we propose a new Fully-Aligned Fusion Network (FAFNet) for RGBD semantic segmentation based on hierarchical semantic flows in this paper. According to the two-modality nature, we employ a two-stream framework as many recent state-of-the-art RGBD segmentation networks[13]. To solve the misalignment issue, a Flow Alignment Module (FAM) is first proposed, that employs a semantic flow to align color and depth pixels at higher precision. For deeper cross-modality feature fusion, we also designed an attention-based module called Attention Cross Fusion Module (ACFM). The channel-attention mechanism can dynamically adjust the weights of two modalities to fit each other and use a cross fusion method to extract more effective feature information. Finally, we packed the FAM and ACFM into one block called RGBDFuse, and then applied it with a multi-layer structure in the FAFNet, which not only enhances the effectiveness of our FAFNet but also overcomes the low resolution and noise drawbacks.

Our contributions can be summarized as followed:

- A FAFNet model for RGBD semantic segmentation is introduced, a hierarchical alignment and fusion structure is designed based on the widely-used two-stream framework*.
- A new RGBDFuse network block for color and depth fusion is proposed, it integrates an attention cross fusion module and two semantic flow modules, which overcomes the cross-modality and misalignment challenges.
- The segmentation accuracy of the proposed method has been evaluated on both the NYU-Depth V2 and the SUN RGB-D benchmark, and comparison with state-of-the-art methods has also been done and presented, which demonstrates the validity of our method.

**The source codes are available at xxx [the link will be inserted after the acceptance of the paper]*

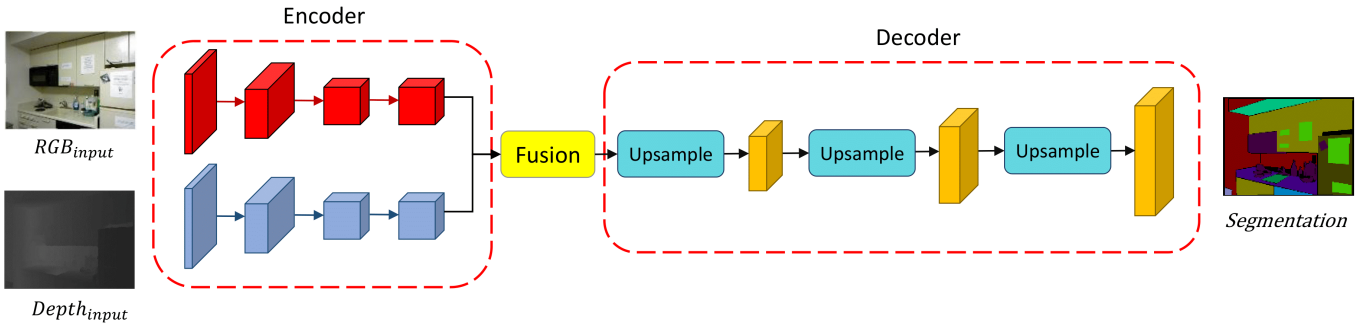


Fig. 1: Overview of existing two-stream neural network framework for RGBD semantic segmentation. Two modalities are processed by two-stream backbone and fused at the end of encoder. This classical method does not recognize the difference between two modalities.

2 Related Work

2.1 RGB Semantic Segmentation

The pioneer of semantic segmentation for natural color images is Fully Convolutional Network (FCN)[1]. It replaces the fully connected layer with a deconvolutional layer with upsampling to achieve pixel-level classification but still suffers from the problem that the upsampled images are not fine enough. U-net[17] uses asymmetric encoder-decoder structure to progressively upsample images to their original size while concatenating high and low-level features together through skip connection to improve accuracy. The Deeplab series[3, 18–20] have introduced atrous convolution to expand the receptive field of convolution kernels and capture contextual information efficiently while keeping parameters constant. RefineNet[21] makes use of the feature maps of each layer, which makes the segmentation more accurate. It also uses residual connection inside the RefineNet Block to make the loss easier to propagate.

Our model in this work is based on the framework of FCN with an encoder-decoder structure. And in the decoder part, high and low-level features are aligned using the FAM to progressively recover the resolution.

2.2 RGBD Semantic Segmentation

A natural manner to integrate the depth map is using depth information as the fourth channel input, as earlier RGBD networks did[22]. However, the results are not effective, due to the different modalities. Inspired by[23], Wang et al. proposed a two-stream network framework to extract specific information from color and depth. A feature transformation module in the middle of the network is designed to explicitly extract common and unique features from color and depth features through the fully connected layer. And then the unique and common features of both are concatenated and fused with a two-stream fully connected layer to the deconvolution layer. Hazirbas et al.[11] proposed a similar two-stream network (FuseNet) framework based on SegNet[24]. They added mid-level depth features to mid-level color features in the encoder by an element-wise addition, and fused the deeper depth features into the color features with deeper network layers. Wang et al. proposed a depth-aware CNN that added a depth similarity term to the normal convolution, and adjusted the weight of surrounding pixels contributing to the central pixels by the similarity of depth values during the convolution operation[25]. Chen et al. proposed a two-stream cross-modality network in which color and depth features were calibrated to remove noises. Then these features are aggregated by attention and gating mechanisms to obtain two weighted feature maps and added together in the final segmentation[13].

The aforementioned networks have certain limitations, they do not recognize the differences between color and depth images. Color and depth images belong to different modalities, and simple concatenating or summing cannot make full use of the complementarity among multiple modalities. In many cases, these two different modalities may suppress each other if they are not well aligned. In this paper, color and depth features are aligned by multi-layer semantic flows modules to get the complementary features, which helps to build an obvious correlation between the two modalities.

2.3 Flow Alignment

The misalignment issue requires image registration between depth and color images, it is a process of transforming different sets of data into one unique coordinate system[26]. Optical flow is a widely-used image registration method, it is originally proposed in motion detection to describe the motion of an observed target, surface, or edge caused by motion relative to the observer. In recent years, optical flow has been used more and more extensively in computer vision-related tasks[27]. Semantic flow proposed by Li et al. extends the concept of optical flow to semantic segmentation[4]. A FAM is proposed to predict the semantic flow between neighboring layer feature maps and efficiently propagate and align high-level features into high-resolution features.

Using semantic flow, semantic features can flow effectively from higher layers to lower layers to reduce the information loss caused by semantic misalignment, but it relies too much on the ability of the backbone to extract features. Offset errors in the backbone tend to be enlarged in the cross-modality feature alignment, which may lead to invalid segmentation results. For this sake, we extended the FAM module to an RGBDFuse block by integrating ACFM. The RGBDFuse block aligns and cross-fuses color and depth features. It not only significantly reduces offset errors but also enhances the feature extraction ability of the backbone.

Existing feature alignment approaches only align single modality features. For instance, classic optical flow methods align pixels between adjacent frames in videos, the feature alignment employed in learning-based object detection aligns the position of features and anchors[28], and Li et al.[4] and Huang et al.[29] proposes flow alignment networks that align high-level and low-level image features. In contrast, our FAFNet further aligns the features between different modalities (i.e. RGB and depth), it extends the feature alignment to the full RGBD semantic segmentation network, thus achieves higher accuracy.

3 Methodology

3.1 The FAFNet structure

RGBD semantic segmentation needs to extract features from both color and depth images separately and then fuse them, which is very challenging due to their different modalities. Fig.1 shows a classical two-stream network framework for RGBD semantic segmentation[1]. In the encoder, color and depth features are extracted separately in a two-stream backbone and then fused together in the end. In the decoder, the fused features are upsampled by using bilinear or de-convolutional layers to recover resolution and obtain the final segmentation results. Though the two-stream network outperforms a single-stream method, it still does not help

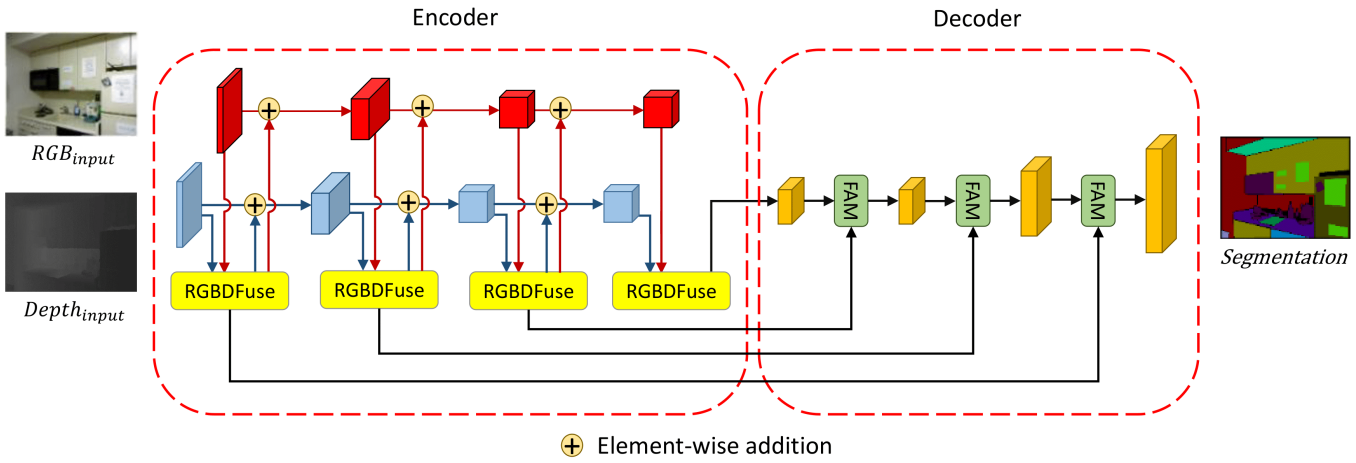


Fig. 2: Overview of our FAFNet. It is based on the encoder-decoder structure. The inputs of our network are two images representing color and depth respectively. Each pair of images is processed by RGBDFuse block and sent back to the backbone to fuse with original color and depth features. The copy of each RGBDFuse block will be propagated to the corresponding layer of decoder to align with high-level features.

existing methods overcome the misalignment issues from either the different modalities or the upsampling process, as introduced in Section 1.

In this paper, we propose a new neural network for RGBD semantic segmentation. The advantage of this network is the fully-aligned feature fusion. As shown in Fig.2, a new cross-modality fusion block called RGBDFuse is proposed. It cross-fuses color and depth features and aligns them into a unified feature map. And a semantic flow module called FAM is integrated into both the RGBDFuse block and the decoder part of our FAFNet, which significantly eases the misalignment issue of RGBD images.

Another important advantage of our FAFNet is the multi-layer fusion structure, compared to the single fusion module in the classical two-stream networks. In our multi-layer structure, multiple RGBDFuse blocks in the encoder and multiple FAMs in the decoder are designed. For each layer in the encoder, color and depth features are sent to the RGBDFuse block for fusion, instead of the next layer of backbone. The RGBDFuse block produces two cross-fused features, which are sent back to the backbone to fuse with original color and depth features respectively. The main feature is kept as a low-level feature in the decoder. For each layer in the decoder, FAM aligns the high-level feature from the previous layer and the low-level feature from the corresponding layer in the encoder. In the end, aligned features are sent to the subsequent layer of the decoder.

Compared to the classical two-stream network, our multi-layer structure aligns not only color and depth features but also high-level and low-level features. It overcomes semantic feature misalignment caused by convolution, upsampling, downsampling, and residual connection in the intermediate layers[4]. No matter in which layer RGBD or high/low-level features become misaligned, they will be re-aligned to provide fully-aligned cross-modality features. Together with the RGBDFuse and FAM, we regard our FAFNet as a fully-aligned fusion network.

3.2 The RGBDFuse block

The purpose of our RGBDFuse block is to fuse cross-modality color and depth features (RGB_{feat} and $Depth_{feat}$ respectively) into one feature $Feat_{fused}$. As shown in Fig.3, it consists of an ACFM module and two FAM modules. The ACFM module employs a cross fusion method to adjust weights for both color and depth features. The FAM modules align these two features through a warping operation with semantic flows. We note these two warped features as RGB_{warped} and $Depth_{warped}$. Finally, an element-wise addition is used to fuse the warped features to obtain fusion features $Feat_{fused}$ containing rich semantic information:

$$Feat_{fused} = RGB_{warped} \oplus Depth_{warped} \tag{1}$$

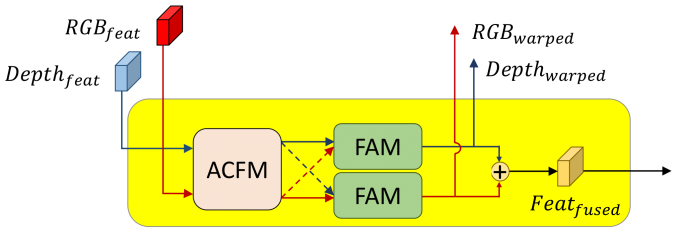


Fig. 3: The RGBDFuse block. An ACFM and two FAMs are employed to fuse complementary features and align two modalities.

Besides, RGB_{warped} and $Depth_{warped}$ are added back to the backbone features and propagated to the subsequent layers, which gets sufficient features for the higher levels:

$$RGB \leftarrow RGB \oplus RGB_{warped} \tag{2}$$

$$Depth \leftarrow Depth \oplus Depth_{warped} \tag{3}$$

In contrast with traditional channel concatenation, our RGBDFuse block explicitly builds the feature correspondences among different modalities based on the attention vectors of concatenated features, which achieves the re-calibration of cross-modality feature weights. The re-calibrated weights can emphasize important features while suppressing redundancy features simultaneously. Besides, features from different modalities are highly complementary, our RGBDFuse block employs a cross-fusion mechanism (i.e. the bidirectional element addition) to deeply fuse them, which further improves the accuracy of feature cross-fusion. More details of ACFM and FAM will be introduced in the two subsections below.

3.3 The ACFM module

To enhance the feature extraction and alignment, we design an ACFM to obtain fused features with richer semantic information. Color and depth features are fed into the ACFM before alignment. As shown in Fig.4, global channel descriptors G_{rgb}/G_{depth} are first extracted. Then attention vectors L_{rgb}/L_{depth} are further squeezed by Multilayer Perceptron (MLP) and Sigmoid functions, and merged with the input color/depth features by channel-wise multiplication. Merged color and depth features (i.e. M_{rgb} and M_{depth}) are cross-fused with each other's original input to obtain rich complementary information. In the remainder of this subsection, we will introduce all of these important steps.

Firstly, the global descriptor is obtained using global average pooling[30]:

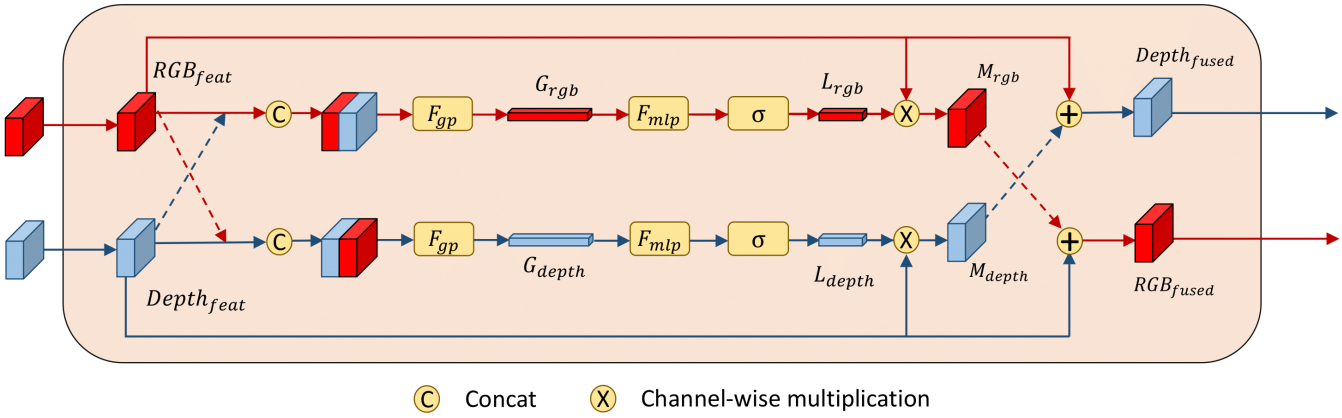


Fig. 4: The ACFM module. Color and depth features are mergerd by global descriptor to adjust the weight of different channels separately, and fused by cross fusion method to aggregate their complementary features into unified features.

$$G = F_{gp}(RGB_{feat} \odot Depth_{feat}), \quad G \in \mathbb{R}^{1\times1\times2C} \tag{4}$$

where G is the global channel descriptor, F_{gp} denotes the global average pooling, C represents the channel number of features, and $\odot$ represents the channel concatenating operation. We obtained G_{rgb} and G_{depth} for both color and depth features respectively.

Secondly, this global channel descriptor is then processed by MLP[31] followed by a Sigmoid normalization operation to obtain an attention vector L:

$$L = \sigma(F_{mlp}(G)), \quad L \in \mathbb{R}^{1\times1\times C} \tag{5}$$

where F_{mlp} is multi-layer perceptron and σ represents the Sigmoid function. The purpose of MLP and Sigmoid function is to extract channel weight information. In our network, a fully connected layer is used to implement MLP. It squeezes the channel number of the global channel descriptor, which allows the network to build the correlation between channels of different modalities and use this correlation to automatically adjust the channel weights for the current modality. Sigmoid function scales the channel weight of MLP's output to $[0, 1]$.

We then fuse the attention vector with one of the original inputs in a channel-wise multiplication. Taking the depth feature as an example, we multiply the attention vector L_{depth} with depth feature input $Depth_{feat}$ to obtain the merged depth feature M_{depth}:

$$M_{depth} = Depth_{feat} \otimes L_{depth}, \quad M_{depth} \in \mathbb{R}^{W\times H\times C} \tag{6}$$

where W and H are the width and the height of the feature map. In the same way, we get merged color features M_{rgb}:

$$M_{rgb} = RGB_{feat} \otimes L_{rgb}, \quad M_{rgb} \in \mathbb{R}^{W\times H\times C} \tag{7}$$

The merged features boost important channel features while suppressing redundancy ones according to the characteristics of the current modality. The boosted features contain features unique to the current modality, which is complementary to the other modality.

Thirdly, a cross fusion is applied to combine the merged features, along with the original input of other modalities. Taking the depth features as an example, M_{depth} is added with RGB_{feat} in an element-wise manner to fuse two modalities' features:

$$Depth_{fused} = M_{depth} \oplus RGB_{feat}, \; Depth_{fused} \in \mathbb{R}^{W\times H\times C} \tag{8}$$

In the same way, we get RGB_{fused} features:

$$RGB_{fused} = M_{rgb} \oplus Depth_{feat}, \; RGB_{fused} \in \mathbb{R}^{W\times H\times C} \tag{9}$$

The cross fusion method fuses merged feature and another modality's original input, which aggregates two modalities' complementary features into unified features.

3.4 The FAM module

For the alignment of high-level and low-level features, most existing neural networks use the bilinear interpolation method for feature upsampling[32]. To recover pixels of the downsampled image, four neighboring pixels are interpolated in a bilinear manner again. However, recovering high-resolution features only using lower-resolution inputs is an under-constraint problem, it tends to result in misalignment between feature maps. As pointed out by Li et al.[4], position correspondence between feature maps needs to be explicitly and dynamically established to resolve the actual misalignment.

For color and depth features, we argue that the semantic flow features can be aligned not only between high-level and low-level features, but also between different modalities. The misalignment of different modalities can cause features to be shifted in the fusion process, resulting in inaccurate feature transfer and making the fusion much less effective. Aligning color and depth modalities can fully expose their unique features.

By integrating multiple FAM densely into our FAFNet, we extend flow alignment to full alignment, which reduces the misalignment issue between different modalities and avoids the feature uneven issue. Whenever RGBD or high-level features become misaligned due to operations such as convolution, upsampling, downsampling, and residual connection, the FAM will recalibrate the features so that they can be well aligned all the times.

In our FAFNet, FAM is used in both the RGBDFuse block and the decoder. FAM in the RGBDFuse block aligns color and depth features from the previous layer. And FAM in the decoder aligns features produced by the RGBDFuse and features from the previous layer in the decoder. Since they share the same structure, we will introduce the detail of this structure below.

As shown in Fig.5, given two feature maps X_i, X_j, we firstly process them with a 1×1 convolution to unify their numbers of channels to have X_i' and X_j'. And then we use bilinear interpolation to upsample the lower resolution feature map X_j' to X_j'' to match the size of X_i'. If X_i and X_j represents RGB_{fused} and $Depth_{fused}$ features respectively in the ACFM module, we skip this upsampling operation since their sizes are the same. Then these two feature maps are concatenated together using the channel concatenating operation and sent to the flow extraction layer to extract the semantic flow feature X_{flow} through a convolution process:

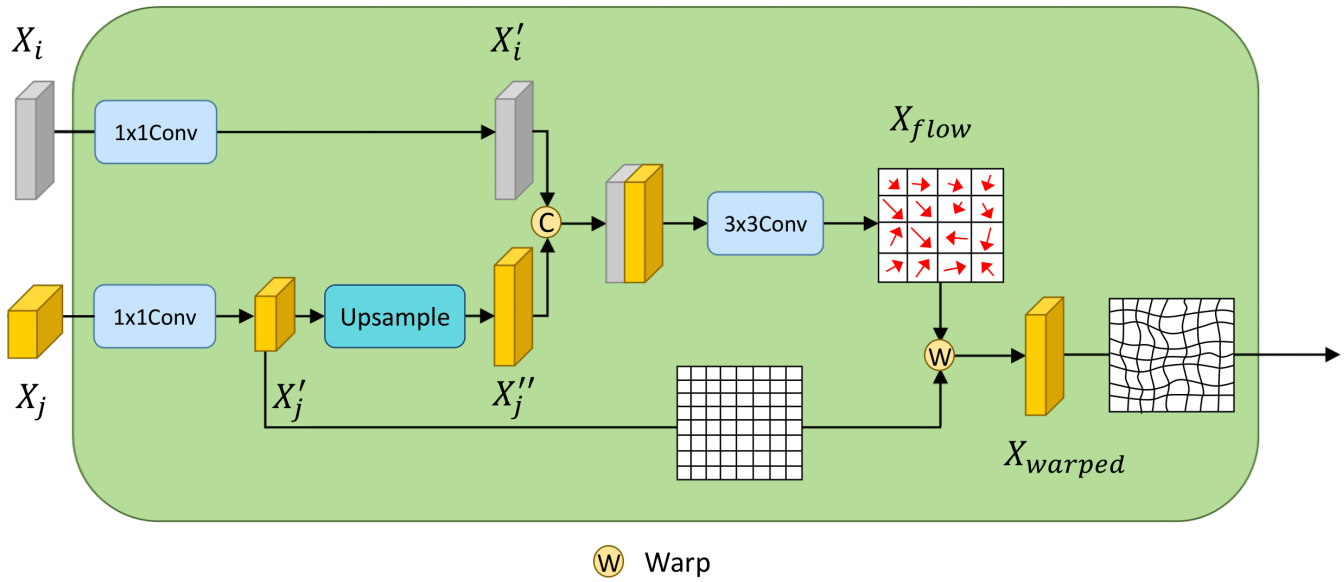


Fig. 5: The FAM module. Two input feature maps are combined to generate a flow feature to align high-level and low-level features as well as color and depth features.

$$X_{flow} = Conv_{flow}(X_i' \odot X_j''), \quad X_{flow} \in \mathbb{R}^{W \times H \times 2} \tag{10}$$

X_{flow} has two channels that represent the offset of pixels in X and Y directions respectively. Based on X_{flow}, positions of X_j' can be shifted to the corresponding positions of X_i' through a warping operation.

Firstly, a spatial grid $\Omega \in \mathbb{R}^{W \times H \times 2}$ is defined, and the number of channels is the same as the channel number of X_{flow}. The size of this spatial grid is the same as the size of the feature map X_i'. Values in each channel are the spatial coordinates corresponding to the input feature map, and their values are divided into normalized coordinates distributed at uniform intervals according to W and H. Let u, v be the coordinate positions of the spatial grid Ω, where u is the integer coordinate in the X-direction and v is the integer coordinate in the Y-direction. Then the value of each feature point on the spatial grid is converted as:

$$\Omega(u, v) = (\frac{2u - W + 1}{W - 1}, \frac{2v - H + 1}{H - 1}) \tag{11}$$

where $u \in [0, W)$ and $v \in [0, H)$. Then, we add the calculated X_{flow} with the spatial grid Ω to obtain a new spatial grid Ω'. The formula is:

$$\Omega' = \Omega \oplus X_{flow}, \quad \Omega' \in \mathbb{R}^{W \times H \times 2} \tag{12}$$

Due to the addition of the semantic flow field offsets, the value of each feature point in Ω' represents the final feature position of the input feature map. According to the values corresponding to the coordinates in Ω', the values of the corresponding feature points are found from the corresponding coordinate positions of the input feature map. The final output high-resolution feature map X_{warped} is finally generated:

$$X_{warped} = Warp(X_j', \Omega') \tag{13}$$

Here the warping operation is used to compute the final feature map by differential image sampling[33]. This warping operation makes it possible to accurately align the low-resolution high-level feature maps to the corresponding positions of the high-resolution low-level feature maps, as well as align different modalities, thus reducing information loss and enhancing the feature extraction capability of the proposed network.

4 Evaluations and Discussion

4.1 Dataset

We evaluate our FAFNet on two widely-used RGBD datasets: NYU-Depth V2[34] and SUN RGB-D[35]. The NYU-Depth V2 dataset consists of many video sequences of various indoor scenes recorded by a Microsoft Kinect V1. It has 1449 RGBD images in 40 classes that are all densely labeled, which are available as GroundTruth (GT) in this database for the training and testing of our FAFNet. We take 795 images for training and 654 for testing. The SUN RGB-D dataset consists of 10,335 RGB-D images in 37 classes that are all densely labeled, and available as the second GT. They are captured by 4 different sensors, including Intel RealSence, Asus Xtion, Microsoft Kinect V1/V2, thus have different image sizes. These devices acquire both RGB images by color cameras and corresponding depth images by depth sensors, such as lasers and infrared cameras for Kinect V1/V2.

We take 5285 images for training and 5050 for testing. Though SUN RGB-D dataset has more images than the NYU-Depth V2 dataset, images in NYU-Depth V2 dataset are annotated more carefully. Therefore, the statistics on the NYU-Depth V2 dataset usually better reveals the accuracy of RGBD semantic segmentation, but the SUN RGB-D dataset is still widely used to evaluate the robustness and generalization ability.

As a preprocessing process, we use random scaling, random cropping, horizontal flipping, and normalization to enhance the dataset. For NYU-Depth V2 dataset, we initialize image size to 464×464. For SUN RGB-D dataset, we initialize image size to 448×448. The single-channel depth image is converted to a three-channel HHA image. The HHA image contains the horizontal disparity, the height to the ground, and the surface normal vector angle, thus makes better use of the depth information. These attributes are hard for the network to compute directly from the depth image, which enhances the geometric structural information in the depth image. For more details of HHA images please refer to[36].

4.2 Implementations

We implemented our FAFNet using the Pytorch 1.8.1 framework, in Ubuntu 18.04 operating system, with Intel 10900KF CPU, NVidia GeForces RTX 3090 graphics card, and 32G RAM.

We use the Mean Intersection over Union (MIoU) to measure the semantic segmentation accuracy. In the training stage, the pre-trained ResNet101[37] is used as the backbone, the learning rate is set to 0.005, and the scheduler strategy we adopted is Step Decay+Warmup[38]. The batch size is set to 8, and the number of iterations is 40,000 in total. We use a multi-scale strategy to boost the performance of our network in the inference stage.

4.3 Ablation study

Uniform hyper-parameters are used to explore which module has the greatest impact on the overall segmentation accuracy. We choose the two-stream Deeplabv3Plus[3] as the baseline, where color and depth features are fused in element-wise summation.

Table 1 Statistics of ablation study for FAM modules on the NYU-Depth V2 Dataset

Method	MIoU	Pix.Acc
Without any FAM	52.3%	77.4%
With FAM in decoder	52.8%	77.7%
With FAM in RGBDFuse	53.1%	78.1%
With all FAMs	**54.0%**	**78.3%**

Table 1 shows the experimental statistics of our ablation study on FAM modules, which demonstrates the effectiveness of our fully-aligned fusion network. Without any FAM, the MIoU is only 52.3%, and the pixel accuracy is only 77.4%. If both the FAM in our RGBD-Fuse and the FAM in the decoder part are used, the MIoU reaches 54.0%, while the pixel accuracy reaches 78.33%. And if only one of these two FAMs is used, it still improves the MIoU and the pixel accuracy, but it does not exceed ones with all FAMs. The ablation study on the ACFM modules has been applied as well. Removing the ACFM module will lead to about 1% drop for both the MIoU and the pixel accuracy, even though all FAMs are used. It shows the ACFM module also contributes to the improvement of our method, as it improves the feature extraction of the backbone.

The fully-aligned fusion network is the key of our method to overcome the misalignment issue of RGB and depth. Fig.6 shows such an example from the NYU-Depth V2 dataset. Since RGB image and depth image are captured by different sensors, they contains unavoidable misalignment even though they are registered, as shown in the RGB-depth overlay in Fig.6(c). The semantic flow estimated in our network shows the depth pixels around the boundary of the kitchen table should be shifted to the left to align with the RGB image, as shown in Fig.6(d). We further extract feature maps of both RGB and HHA images from the ACFM module of the first RGBDFuse block, to compare their alignment before and after cross-modality feature alignment. The overlay of RGB and HHA features in Fig.6(f) is distinctly better aligned than the overlay in Fig.6(e), which reveals the validity of our cross-modality feature fusion network.

4.4 Comparison

Table 2 Comparison with SOTA methods on the NYU-Depth V2 Dataset

Method	Backbone	Data	MIoU	Pix.Acc
ShapeConv[39]	ResNeXt-101×2	RGB-HHA	51.3%	76.4%
ACNet[40]	ResNet-50×2	RGB-Depth	48.3%	-
RDFNet[41]	ResNet-152×2	RGB-HHA	50.1%	76.0%
CANet[12]	ResNet-101×2	RGB-Depth	51.2%	76.6%
NANet[42]	ResNet-101×2	RGB-HHA	52.3%	77.9%
SA-Gate[13]	ResNet-101×2	RGB-HHA	52.4%	77.9%
Ours	ResNet-101×2	RGB-HHA	**54.0%**	**78.3%**

Table 2 shows a quantitative comparison with state-of-the-art (SOTA) methods on the NYU-Depth V2 dataset. Both the MIoU and the pixel accuracy of our method outperform other SOTA networks, the MIoU of our method exceeds SA-Gate about 1.6%, the pixel accuracy of our method exceeds SA-Gate about 0.4%.

Table 3 Comparison with SOTA methods on the SUN RGB-D Dataset

Method	Backbone	Data	MIoU	Pix.Acc
ShapeConv[39]	ResNeXt-101×2	RGB-HHA	48.6%	82.2%
ACNet[40]	ResNet-50×2	RGB-Depth	48.1%	-
RDFNet[41]	ResNet-152×2	RGB-HHA	47.7%	81.5%
CANet[12]	ResNet-101×2	RGB-Depth	49.3%	**82.5%**
NANet[42]	ResNet-101×2	RGB-HHA	48.8%	82.3%
SA-Gate[13]	ResNet-101×2	RGB-HHA	**49.4%**	**82.5%**
Ours	ResNet-101×2	RGB-HHA	49.2%	82.3%

Table 3 shows a quantitative comparison with state-of-the-art methods on the SUN RGB-D dataset. Our FAFNet is at the forefront of state-of-the-art methods, only 0.2% weaker than SA-Gate, reflecting the robustness and generalization ability of our network. The main reason why we did not achieve the best results on the SUN RGB-D dataset is that there are a large number of images with different resolutions in SUN RGB-D, and padding these images affects the accuracy of the flow field. In addition, it can be seen from the table that most existing work also rarely achieves the best performance on both NYU-Depth V2 and SUN RGB-D datasets.

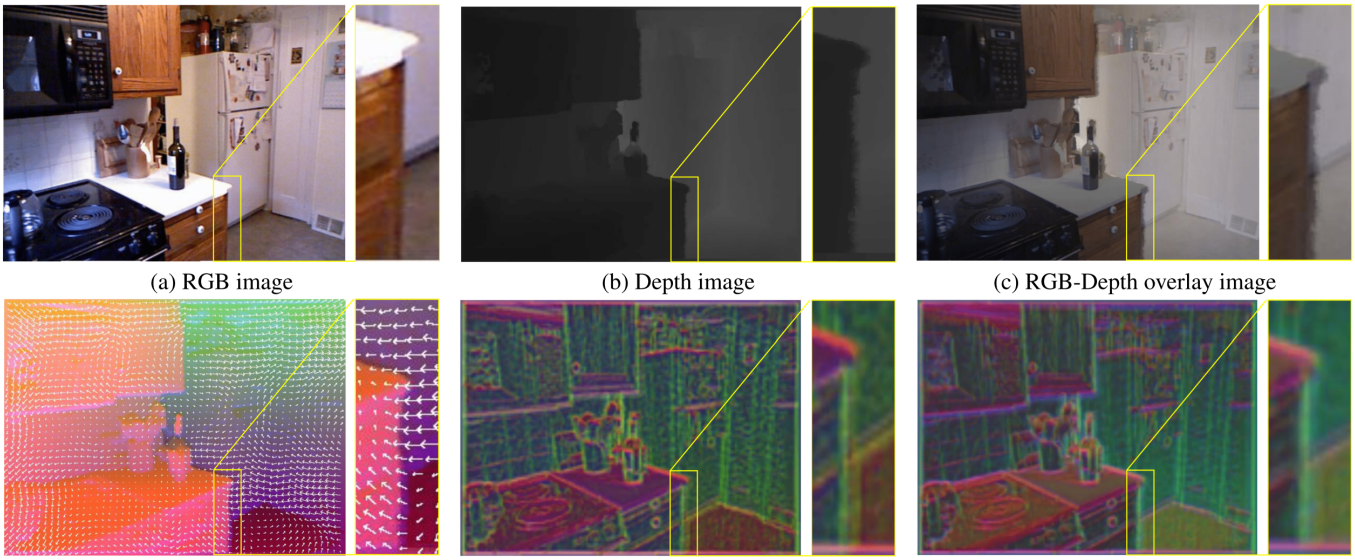
Table.4 shows the class-wise segmentation accuracy of our network compared with SA-Gate[13]. It can be seen that our network shows significant improvement in most classes, 29 out of 40 classes exceeded SA-Gate, especially in the class bag, which shows a 7% improvement. The average growth of these 29 classes is 2.78%, in contrast, the average decrease of the rest 11 classes is only 1.5%. Some of the classes that are prone to misalignment in the depth map, such as $refridge$, tv, and $paper$, our method obtains about 3-4% improvement in average.

Fig.7 shows a qualitative comparison of our method with SA-Gate[13]. To visualize the segmentation results, we applied a pseude coloring to the regions of differnt categories, as well as the ground truth. It is very challenging to segment the black chair in the first row, even though the HHA image provides supplementary geometric information. It can be seen that our FAFNet distinguishes the contour of the chair very well, while SA-Gate mixes the chair with the floor. The floor lamp in the second row is also difficult to distinguish as well, SA-Gate lost the lamp pillar while our method succeeded to preserve this pillar. There are obvious errors for the SA-Gate segmentation of the bedsheet and bedside table in the third row, while our FAFNet segments them much better, indicating that our network has strong feature extraction capabilities.

5 Conclusion and Future Work

In this paper, we present a new two-stream neural network for RGBD semantic segmentation called FAFNet, which can fully align color and depth features. A cross-modality feature alignment and fusion block called RGBDFuse is proposed, it integrates two semantic flows and an attention cross fusion model to overcome the RGBD misalignment issue and the cross-modality challenge. A hierarchical structure, that applies the RGBDFuse block and flow alignment modules for features at multiple layers, is employed to further ease the resolution and noise difference between depth and color images. Using the NYU-Depth V2 and SUN RGB-D dataset, evaluations including ablation studies and elaborated comparison show that the proposed FAFNet achieves higher accuracy than state-of-the-art RGBD semantic segmentation.

One limitation of our method is that our FAFNet does not manage to segment small objects well, this drawback can also be found in many other RGBD semantic segmentation methods. We believe more improvement could be made to recognize small objects in the



(a) RGB image (b) Depth image (c) RGB-Depth overlay image

(d) Semantic flow visualization using arrows (e) RGB-HHA feature overlay without alignment (f) RGB-HHA feature overlay with alignment

Fig. 6: Visualization of our feature alignment results on one example from NYU-Depth V2 dataset. The overlay of RGB and depth images clearly reveals their misalignment in (c). The estimated semantic flow is visualized using colors and arrows in (d). The direction of white arrows shows the flow direction, while the length of the white arrows reveals the strength of semantic flow at small grid. Feature maps for RGB and depth are overlayed to show their misalignment in (e), which is alleviated if semantic flow is employed in (f). The zoom-in insets on the right emphasize how our semantic flow eases the misalignment issue on the boundary of the kitchen table.

Table 4 Class-wise segmentation results (MIoU) on NYU-Depth V2 Dataset

	wall	floor	cabinet	bed	chair	sofa	table	door	window	bkshelf
SA-Gate	80.8%	88.6%	**64.4%**	**75.1%**	67.2%	68.5%	48.5%	**46.6%**	**52.3%**	46.4%
	81.2%	**88.8%**	63.4%	72.9%	**68.6%**	**69.5%**	**50.8%**	46.0%	50.5%	**49.9%**
	picture	counter	blind	desk	shelf	curtain	dresser	pillow	mirror	mat
SA-Gate	66.6%	71.8%	**60.5%**	26.6%	**22.2%**	**60.1%**	54.1%	51.1%	**58.1%**	42.5%
Ours	**66.9%**	**72.4%**	59.2%	**30.5%**	21.1%	59.6%	**55.4%**	**54.8%**	55.3	**42.7%**
	cloths	ceiling	books	refridge	tv	paper	towel	shower	box	board
SA-Gate	24.1%	**80.9%**	30.5%	61.1%	65.3%	32.9%	43.3%	41.2%	13.1%	78.3%
Ours	**26.4%**	80.0%	**34.9%**	**65.2%**	**69.6%**	**35.9%**	**45.8%**	**45.9%**	**15.2%**	**85.0%**
	person	stand	toilet	sink	lamp	bathtub	bag	othstr	othfurn	othprop
SA-Gate	83.2%	41.0%	81.0%	**62.6%**	50.7%	**56.4%**	5.2%	31.2%	21.8%	41.5%
Ours	**87.8%**	**46.9%**	**84.7%**	59.7%	**53.4%**	55.0%	**12.2%**	**32.6%**	**22.8%**	**42.8%**

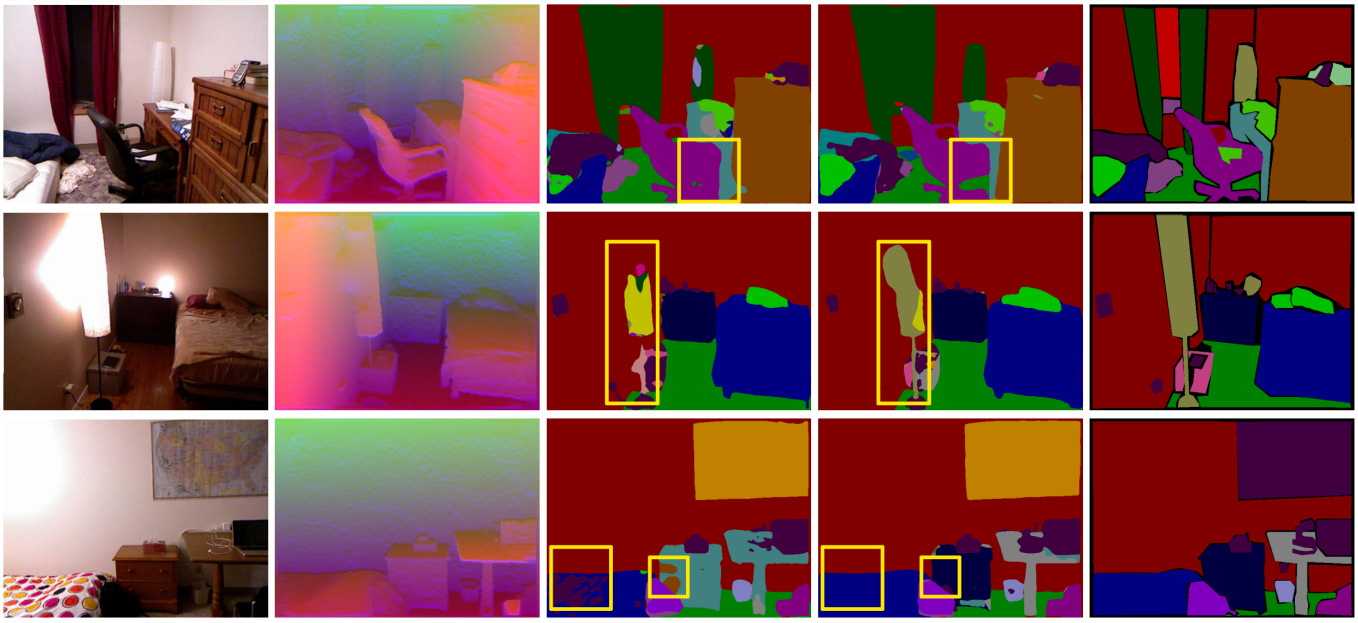


Fig. 7: Comparison with the SOTA method SA-Gate[13] on three examples from the NYU-Depth V2 dataset

upsampling procedure of high-level features. Though our method reaches the highest accuracy on the NYU-Depth V2 dataset, but not on the SUN RGB-D. One future direction is thus to improve the compatibility of our method on more messy datasets. Another future direction is to accelerate our FAFNet to achieve real-time performance, which may require simplifying our FAFNet to a lightweight version.

Acknowledgments

This work was supported by National Natural Science Foundation of China (62172367) and Zhejiang Science and Technology Project of Cultural Relic Protection (2020014).

6 References

1 Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3431–3440, 2015.
2 Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2881–2890, 2017.
3 Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.
4 Xiangtai Li, Ansheng You, Zhen Zhu, Houlong Zhao, Maoke Yang, Kuiyuan Yang, Shaohua Tan, and Yunhai Tong. Semantic flow for fast and accurate scene parsing. In *Proceedings of European Conference on Computer Vision (ECCV)*, pages 775–793. Springer, 2020.
5 Neda Noormohamadi, Peyman Adibi, and Sayyed Mohammad Saeed Ehsani. Semantic image segmentation using an improved hierarchical graphical model. *IET Image Processing*, 12(11):1943–1950, 2018.
6 Yashwant Kurmi and Vijayshri Chaurasia. Multifeature-based medical image segmentation. *IET Image Processing*, 12(8):1491–1498, 2018.
7 Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.
8 Richard Nock and Frank Nielsen. Statistical region merging. *IEEE Transactions on pattern analysis and machine intelligence*, 26(11):1452–1458, 2004.
9 Po Kong Lai, Weizhe Liang, and Robert Laganière. Additive depth maps, a compact approach for shape completion of single view depth maps. *Graphical Models*, 104:101030, 2019.
10 Yilin Liu, Ke Xie, and Hui Huang. Vgf-net: Visual-geometric fusion learning for simultaneous drone navigation and height mapping. *Graphical Models*, 116:101108, 2021.
11 Caner Hazirbas, Lingni Ma, Csaba Domokos, and Daniel Cremers. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture. In *Proceedings of Asian Conference on Computer Vision (ACCV)*, pages 213–228. Springer, 2016.
12 Hao Zhou, Lu Qi, Zhaoliang Wan, Hai Huang, and Xu Yang. Rgb-d co-attention network for semantic segmentation. In *Proceedings of the IEEE Asian Conference on Computer Vision (ACCV)*, 2020.
13 Xiaokang Chen, Kwan-Yee Lin, Jingbo Wang, Wayne Wu, Chen Qian, Hongsheng Li, and Gang Zeng. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 561–577. Springer, 2020.
14 Zhen Li, Yukang Gan, Xiaodan Liang, Yizhou Yu, Hui Cheng, and Liang Lin. Lstm-cf: Unifying context modeling and fusion with lstms for rgb-d scene labeling. In *Proceedings of European conference on computer vision (ECCV)*, pages 541–557. Springer, 2016.
15 Xiaojuan Qi, Renjie Liao, Jiaya Jia, Sanja Fidler, and Raquel Urtasun. 3d graph neural networks for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 5199–5208, 2017.
16 Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.
17 Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.
18 Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*, 2014.
19 Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017.
20 Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 801–818, 2018.
21 Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1925–1934, 2017.
22 Camille Couprie, Clément Farabet, Laurent Najman, and Yann LeCun. Indoor semantic segmentation using depth information. *arXiv preprint arXiv:1301.3572*, 2013.
23 Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 1520–1528, 2015.
24 Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(12):2481–2495, 2017.
25 Weiyue Wang and Ulrich Neumann. Depth-aware cnn for rgb-d segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 135–150, 2018.
26 Anton Bardera, Imma Boada, Miquel Feixas, Jaume Rigau, and Mateu Sbert. Multiresolution image registration based on tree data structures. *Graphical Models*, 73(4):111–126, 2011.
27 Xiuxiu Li, Yanjuan Liu, Haiyan Jin, Jiangbin Zheng, and Lei Cai. Automatic layered rgb-d scene flow estimation with optical flow field constraint. *IET Image Processing*, 14(16):4092–4101, 2020.
28 Jiaming Han, Jian Ding, Jie Li, and Gui-Song Xia. Align deep features for oriented object detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022.
29 Zilong Huang, Yunchao Wei, Xinggang Wang, Wenyu Liu, Thomas S. Huang, and Humphrey Shi. Alignseg: Feature-aligned segmentation networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):550–557, 2022.
30 François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1251–1258, 2017.
31 Matt W Gardner and SR Dorling. Artificial neural networks (the multilayer perceptron)–a review of applications in the atmospheric sciences. *Atmospheric environment*, 32(14-15):2627–2636, 1998.
32 Joshua B Tenenbaum and William T Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–1283, 2000.
33 Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2017Í2025, Cambridge, MA, USA, 2015. MIT Press.
34 Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Proceedings of European conference on computer vision (ECCV)*, pages 746–760. Springer, 2012.
35 Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015.
36 Saurabh Gupta, Ross Girshick, Pablo Arbeláez, and Jitendra Malik. Learning rich features from rgb-d images for object detection and segmentation. In *Proceedings of European conference on computer vision (ECCV)*, pages 345–360. Springer, 2014.
37 Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.
38 Rong Ge, Sham M Kakade, Rahul Kidambi, and Praneeth Netrapalli. The step decay schedule: A near optimal, geometrically decaying learning rate procedure for least squares. *arXiv preprint arXiv:1904.12838*, 2019.
39 Jinming Cao, Hanchao Leng, Dani Lischinski, Daniel Cohen-Or, Changhe Tu, and Yangyan Li. Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7088–7097, 2021.
40 Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pages 1440–1444, 2019.
41 Seong-Jin Park, Ki-Sang Hong, and Seungyong Lee. Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4980–4989, 2017.
42 Guodong Zhang, Jing-Hao Xue, Pengwei Xie, Sifan Yang, and Guijin Wang. Non-local aggregation for rgb-d semantic segmentation. *IEEE Signal Processing Letters*, 28:658–662, 2021.