IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received April 1, 2021, accepted April 15, 2021, date of publication April 28, 2021, date of current version May 12, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3076346

Monocular Depth Estimation Based on Multi-Scale Depth Map Fusion

XIN YANG[1,2], QINGLING CHANG[1,2], XINGLIN LIU[1,2], SIYUAN HE[1,2], AND YAN CUI[1,2,3]

[1]Faculty of Intelligent Manufacturing, Wuyi University, Jiangmen 529000, China
[2]China-Germany Artificial Intelligence Institute (Jiangmen), Jiangmen 529000, China
[3]Zhuhai 4Dage Network Technology, Zhuhai 519000, China

Corresponding author: Yan Cui (cuiyan@wyu.edu.cn)

This work was supported in part by the School Research Projects of Wuyi University under Grant 2019AL032 (Knowledge Graph Construction for Big Data of Science and Technology Resources).

ABSTRACT Monocular depth estimation is a basic task in machine vision. In recent years, the performance of monocular depth estimation has been greatly improved. However, most depth estimation networks are based on a very deep network to extract features that lead to a large amount of information lost. The loss of object information is particularly serious in the encoding and decoding process. This information loss leads to the estimated depth maps lacking object structure detail and have non-clear edges. Especially in a complex indoor environment, which is our research focus in this paper, the consequences of this loss of information are particularly serious. To solve this problem, we propose a Dense feature fusion network that uses a feature pyramid to aggregate various scale features. Furthermore, to improve the fusion effectiveness of decoded object contour information and depth information, we propose an adaptive depth fusion module, which allows the fusion network to fuse various scale depth maps adaptively to increase object information in the predicted depth map. Unlike other work predicting depth maps relying on U-NET architecture, our depth map predicted by fusing multi-scale depth maps. These depth maps have their own characteristics. By fusing them, we can estimate depth maps that not only include accurate depth information but also have rich object contour and structure detail. Experiments indicate that the proposed model can predict depth maps with more object information than other prework, and our model also shows competitive accuracy. Furthermore, compared with other contemporary techniques, our method gets state-of-the-art in edge accuracy on the NYU Depth V2 dataset.

INDEX TERMS Monocular depth estimation, dense feature fusion network, depth adaptive fusion module, multi-scale depth maps, indoor.

I. INTRODUCTION

Depth estimation is a fundamental problem in computer vision, applied to robot navigation, augmented reality, 3D reconstruction, autonomous driving, and other fields. From the last century, scholars have begun to try to estimate the depth of the scene. Previous methods usually use optical and environmental geometric constraints to estimate the depth of the scene [1]. At present, most depth estimation is based on the transformation estimation of two-dimensional RGB images to RBG-D images, mainly including the Shape From X method to obtain the scene depth shape from the image brightness and shade, different perspectives, luminosity and texture information, as well as the algorithm combining SFM (Structure From Motion) and SLAM(Simultaneity Localization And Mapping) methods to predict the camera position and pose. Although many devices can obtain depth directly, the equipment is expensive. Binocular depth estimation can also be used, but to deal with binocular images, it needs to use stereo matching for pixel correspondence and parallax calculation. The computational complexity is very high, and for the low texture scene, the matching effect is not good. Though monocular depth estimation is considered an ill-posed problem for a single image that lacks the necessary geometric information, it has been widely concerned because it's relatively cheaper and easier to popularize. Monocular depth estimation is to use an RGB image under a single or unique perspective to estimate the distance of each pixel in the image relative to the shooting source. With the development of technology, monocular depth based on deep learning has

The associate editor coordinating the review of this manuscript and approving it for publication was Alex Noel Joseph Raj.

made great progress. Eigen *et al.* [2] proved through experiments that the relative depth of the image can be obtained by using a Convolutional Neural Network(CNN). Most of the Newest approaches [3]–[6] about depth estimation based on deep learning show great performance. However, there are some challenges in depth estimation from a single image.

First, most models of depth estimation are based on very deep neural networks to extract the features from the image to get good performance, but the feature maps obtained by multiple convolutions lose many pieces of information especially object information, which leads to small objects and object structure detail missed in the feature map. Furthermore, after many times of convolution and pooling, the spatial dependence and channel dependence of the feature map is weak. We will preserve much object information, but can't obtain high-level semantic information if we reduce the convolutional layers. Generally, most depth information associated with high scale feature maps, and object information are usually associated with lower feature maps. For indoor scenes with many objects, the impact of information loss is particularly serious. To deal with this problem, some pre-works [7], [8] introduce the skip-connection to add Low-scale features to the decoder module. But most of the time, skip-connection only includes a single-scale feature map.

Second, many pre-works reference the U-NET [9] architecture to design the network. Although U-NET gets good performance in many vision tasks, the gradual decoding makes U-NET shows poor performance in multi-scale feature fusion. To deal with those problems, we propose a network to estimate the depth from a single image by fuse multi-scale depth maps. In the part of the encoder, we propose a Dense Feature Fusion Network (DFFN). DFFN uses multiple sampling and channel compression to fuse different scale features for obtaining fused feature blocks preserving the detailed information of the image. Since the features extracted by convolution are a mixture of cross-channel and spatial information that leads to the dependence between features in the feature maps is less, we introduce the channel attention module to enhance dependence. In the decoder part, we designed a Depth Adaptive Fusion Module (DAFM) to decode the feature block. In this module, we estimate the depth maps from each scale feature block and set learnable weights for these depth maps. Finally, we get the final feature map by directly adding these weighted depth maps. Using this module, we not only get a depth map with rich scene information but also reduce the parameters of the decoding module.

Our contributions can be summarized as follows:

- We propose a Dense Feature Fusion Network (DFFN). By aggregating multi-scale features, a feature pyramid is established to solve problems, caused by the loss of information in the encoding process, such as the blur of the depth map and the lack of object structure information.
- We propose the Deep Adaptive Fusion Module (DAFM). By estimating coarse depth maps of various scales, and performing weighted summation on these depth maps, a depth map with both high accuracy and rich scene information is obtained.
- Extensive experimental results show that our model shows that our predicted depth map has more object information and clearer edges than other previous works, and has competitive depth accuracy in the NYU-Depth V2 dataset.

II. RELATED WORK

In this section, we mainly discuss the related works about monocular depth estimation and multi-scale fusion.

A. MONOCULAR DEPTH ESTIMATION

In recent years, deep convolutional networks have been applied to depth estimation and have achieved excellent results such as [2]–[8], [10]–[24]. Now we generally considered that the beginning of the depth estimation of a single image based on deep learning is Eigen *et al*. [2]. Eigen used a multi-scale convolutional neural network to extract image features and predicted image depth. Although they got a coarse depth map, they proved that the depth of an image can be estimated by extracting the multi-scale features of the image. Based on this work, Eigen and Fergus [10] built a framework composed of depth estimation, surface normal prediction, and semantic annotation. Liu *et al.* [11] used a deep convolutional neural network and Conditional Random Field (CRF) to imitate the intricate relationship between adjacent parts of the depth map precisely. The model extracted the relevant features from an RGB image through CNN, then used CRF to improve the smoothness and edge preservation of adjacent superpixel blocks. Laina *et al*. [7] proposed a network based on FCRN (Fully Convolutional Residual Networks) to a depth map or depth maps. Yuru *et al.* [24] added an attention mechanism to the classification algorithm, combined with contextual content, and it also used the soft classification method to improve the quality of prediction depth. Wu *et al.* [23] applied (Atrous Spatial Pyramid Pooling) ASPP to depth estimation tasks. It used ASPP convolution kernels of different sizes to obtain feature information of different scales, which achieved excellent estimation results. Lo *et al*. [22] proposed a multi-channel and multi-rate feature extractor, which can effectively extract multi-scale information for depth prediction. Most of these methods have not shown great ability to recover the information of the objects in the depth map, resulting in the blurry objects in the estimated depth map, especially many structure details are lost. In addition to the above-supervised methods, to reduce reliance on labeled data, scholars have also introduced unsupervised estimation methods [12], [13], [18] that utilized epipolar geometry and deep CNN to train the network. SfMLearner [25] is the first framework to predict depth and ego-motion using monocular videos. In this paper, we focus on the supervised depth estimation method.

B. MULTI-SCALE FEATURE FUSION

Multi-scale feature fusion is a common method of feature extraction. Convolutional neural networks gradually abstract and extract features through convolution. The deep network has a large receptive field and strong semantic expression ability. However, due to the deep layer convolutional abstractions, the resolution of the feature map is low, and a large number of spatial characteristics detail information is lost. The shallow network shows the ideal ability to preserve scene information and the feature maps are higher resolution, but the semantic information representation ability is weak. Multi-scale feature fusion deals with these problems. The fused features from different scales can help us get the fusion features that have good semantic expression ability and many spatial characteristics detail information, such as [3], [26]–[30]. He *et al.* [26] proposed (Spatial Pyramid Pooling in Deep Convolutional Networks) SPPNet, which used atrous convolution with different atrous rates to extract the feature map at the same time, and then the features were all connected by channels to obtain a fused feature map. Zhao *et al.* [27] proposed (Pyramid Scene Parsing Network)PSPNet. They used different sizes pooling layers on the same feature map at the same time by constructing a pooling pyramid to obtain feature maps of different scales. Then they upsampled all the feature maps of different scales to the same resolution as the original feature map. Finally, they connect them to the original features map. Liu *et al.* [28] proposed the ParseNet, which extracts the global features of the image through global pooling and merges them with local features. Zhou *et al.* [29] in order to explore the complex relationships and exploit the complementarity between RGB image and depth information, the Deep Convolutional Residual Autoencoder (DCRA) includes two branch input RGB branch and input depth branch were proposed. Furthermore, multi-fusion modules were proposed at the same time. These modules aggregate information include texture and structure information between the RGB and depth branches of the encoder and fuse their features over several multiscale layers. These modules showed a great performance in the feature fusion. Wu *et al.* [30] proposed a multi-level context and multimodal fusion network (MCMFNet) to fuse multi-scale multi-level context feature maps and learn the object edges from depth information. MCMFNet can obtain the detected result with a clear object boundary. Hu *et al.* [3] proposed a network using convolution instead of the pooling layer for multi-scale feature fusion. they upsampled all the feature maps obtained by the encoder to the same resolution with channel compression, channel connection, and convolution operation. But the directly upsampling high scale feature map just increases the resolution and does not increase additional features. Furthermore, there is a lack of feature dependence between feature maps of different scales, which with impact fuse result. To improve the fusion effectiveness, we downsample the low scale feature and concatenate it with high scale features and then compress them before upsampling the high scale features. Furthermore, we introduce channel attention to assist the feature fusion. Experiments show that our proposed Dense Feature Fusion Network achieved good results based on our baseline. We will introduce the detail in Section III.

III. OVERVIEW

In this section, we first introduce the overall framework, and then we describe the Dense Fuse Feature Network (DFFN) and the Depth Adaptive Fusion Module (DAFM) in detail.

A. OVERALL FRAMEWORK

In this paper, we select the encoder-decoder framework as the base architecture. In the encoding phase, we proposed a Dense Feature Fusion Network (DFFN), that can obtain fused feature maps of multiple scales by aggregating different scales' feature maps. These fused feature maps will participate in decode together with the original feature maps to solve the problem of information loss. In the decoder module, unlike the previous work, we use the Depth Adaptive Fusion Module (DAFM) to fuse the rough depth map estimated from each scale feature block to obtain the final depth map. The framework of our network is shown in FIGURE 1. Specifically, (1) the encoder module uses SENet [31] as the backbone. To deal with the information loss in the deep network, the Dense Feature Fusion Network (DFFN) is proposed. We denote the fusion feature map, on the *ith* layer of the fusion feature pyramid, as F_f^i, and the F_{ori}^i indicating the original feature maps at the *ith* level of the pyramid of the backbone. We will introduce the detail about the fusion network in the next part. (2) In the decoder phase, we proposed the Depth Adaptive Fusion Module (DAFM) to get the fine depth map by fuse multi-scale depth maps. We first compress the original feature map to half of the channels and connect them to obtain feature blocks. Through this operation, we ensure that the coarse depth map estimated in the next operation contains object information and depth information. Furthermore, concatenate fused and original feature maps can close the difference in those depth maps. Next, we compress these feature blocks to get the coarse multi-scale depth maps and set a learn the weight for each coarse depth map. At last, we sum the weighted depth maps to obtain the final depth map.

B. DENSE FEATURE FUSE NETWORK

Generally, high-scale feature maps have rich semantic information but lack spatial geometric features. More importantly, many objects will lose in high scale feature maps for the multi-scale pooling layers, which will lead to a bad result in depth prediction. This issue is severe especially in indoor environments that contain many objects and the scene is more complex than outdoor. To deal with this problem, a dense feature fuse network is proposed and we use a novel preprocessing strategy to the upsampling of low-resolution feature maps in the fusion network and the upsampling method we following [7]. When we upsample the low-resolution feature map, we do not directly perform upsampling, but we first

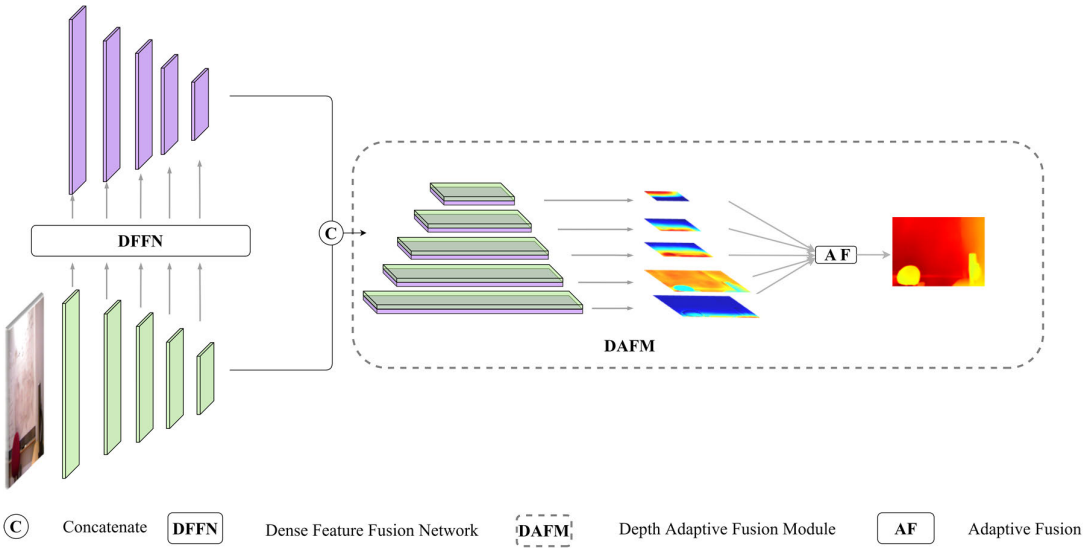


FIGURE 1. The architecture of our network. In our encoder module, we proposed a Dense Feature Fusion Network(DFFN) to fuse a multi-scale feature and build an FFP(Feature Fused Pyramid). In the decoder module, we propose the Depth Adaptive Fusion Module to obtain fine depth map by adaptive fuse multi-scale depth maps.

down-sample the high-resolution feature map of the target, and then we connect it with the low-resolution feature map. Finally, the connected feature block is upsampled. However, if each low-resolution feature map upsample using this strategy, that will add too many redundant features and additional computation. Therefore, in the fusion network, we only use this strategy in perform sampling on non-adjacent scale feature maps. Specifically, we note the *ith* scale feature map upsampling to *kth* scale as $F^{i\to k}i \in \{1, 2, 3\cdots n\}$. We only use the preprocessing strategy when $k - i \geq 2$. We directly upsampling the feature map when $k-i \leq 2$. Then we concatenate $\{F^{i\to k}\}$ that the k is the same as a feature map. Through this strategy, we can ensure that the low-fused map includes more object information. Finally, we use a channel attention mechanism to enhance the dependence of these concatenated feature maps and express it as F_f^k. We show the fusion submodule in DFFN to create F_f^1 in FIGURE 2 (a). The module structure of channel attention is shown in FIGURE 2 (b). In our attention mechanism module, we use average pooling to shape the concatenated fused map to become C $\times$ 1 $\times$ 1. Then, we utilize a 1 $\times$ 1 convolution operation to reduce their number of channels to half F_{ori}^k (F_{ori}^k indicating the original feature maps at the *kth* level of the pyramid of the backbone) channels and use Relu to activate. After that, we use 1 $\times$ 1 convolution again but without changing the channels and use sigmoid as activate function and express it as att_k. On the other hand, we use the 1 $\times$ 1 convolution to compress F_{ori}^k to half channels and multiply it with att_k. Finally, we use a 3×3 convolution to process the multiplied feature map to get the fused feature F_f^k.

C. ADAPTIVE DEPTH FUSION MODULE

In this section, we propose an adaptive depth fusion module. As we know most of the traditional decoding module uses a step-by-step decoding method and the feature map generated by the previous decoding will be passed up and participate in the next step of decoding [2], [3], [19], [22]. The step-by-step not only increase the parameters of the decoding module but also show a low ability in multi-scale feature fusion. Different scale depth maps have different characteristics. High scale deep maps show a good performance in depth information prediction, but there is almost without spatial information and object information of the scene, whether object position or object outline. Rich scene information and detailed information exist in the low-scale depth map but hardly include any depth information. The comparison of different scale depth maps is shown in FIGURE 3. As we can see in FIGURE 3 that the 1th and 2th scale depth map has rich spatial information but they have not depth information or the depth information is inaccurate. Furthermore, there is depth information and without any object information in the 5th scale depth map. If we can fuse different scale depth maps, we can get a depth map including depth value and scene and objects information through fuse depth maps of multi-scale. Inspired by [32], we proposed an adaptive depth fusion module to fuse multi-scale depth maps by weighted summation. The architecture is shown in FIGURE 4. First, we perform channel compression on all the original feature maps and connect the compressed feature maps with the fused feature maps of the corresponding scale. By connecting the fusion feature maps which include all scale features with the original feature map of the corresponding scale, the information gaps between the depth maps of different scales are reduced and improve the fusion effect. If we use the original feature map alone, the fusion effect will be impacted because the features between different scales have too large gaps. The corresponding ablation experiment is in Sec IV. Through these operations, we will get multiple feature blocks. Then we perform channel compression on each feature block

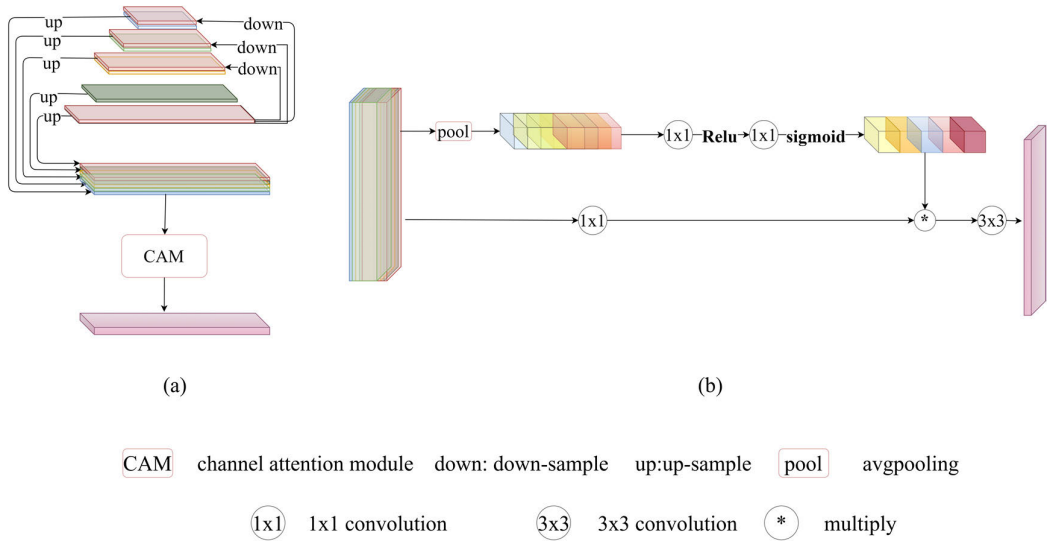


FIGURE 2. The architecture of the submodule and attention module in our dense feature network. (a) Our submodule was used to calculate the first-scale fused feature map in DFFN. DFFN includes 5 submodules similar to (a). These submodules are used to fuse the feature maps of scales 1-5. (b) Our channel attention module.

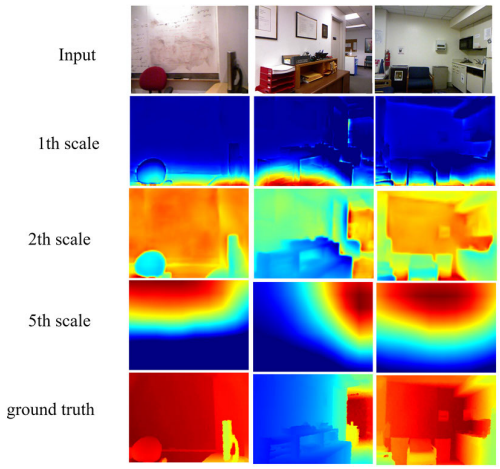


FIGURE 3. Different scale depth map. From top to bottom are the original image, the first-scale depth map, the second-scale depth map, the fifth-scale depth map, and the ground truth. First-scale and second-scale depth maps include rich object information. The fifth-scale depth map includes depth information.

separately to obtain a rough depth map. To improve the presentation ability of the network, we adopted a stepwise compression strategy when compressing the feature map to predict coarse depth maps. We use the 3×3 convolution to compress their channel to half and use batch norm and Relu to deal with these feature maps. After these operations, we use the 3×3 convolution to compress the feature map to a single channel. Finally, we use bilinear interpolation to resize the multi-course depth maps to the same size and set a learnable weight w_k for each depth map d_k. The final depth map D_f can be expressed as (1).

$$D_f = \sum_{k \in l} w_k * d_k \tag{1}$$

where n is the number of the scales, the k express the kth scale, $l = [1, 2, 3 \cdots n]$. In this paper, we set $n = 5$. D_f is the final depth map. The depth value in (i, j) can be denoted as (2)

$$Y_{i,j} = \alpha_{i,j} y^1_{i,j} + \beta_{i,j} y^2_{i,j} + \gamma_{i,j} y^3_{i,j} + \delta_{i,j} y^4_{i,j} + \varepsilon_{i,j} y^5_{i,j} \tag{2}$$

where $y^k_{i,j}$ is the depth value of (i, j) in d_k, $\alpha_{i,j}\beta_{i,j}\gamma_{i,j}\delta_{i,j}\varepsilon_{i,j}$ is the weight of $y^k_{i,j} k \in \{1, 2, 3, 4, 5\}$. Furthermore, we set two mathematical constraints to the weight as (3) and (4)

$$\alpha_{i,j}, \beta_{i,j}, \gamma_{i,j}, \delta_{i,j}, \varepsilon_{i,j} \subset [0, 1] \tag{3}$$

And

$$\alpha_{i,j} + \beta_{i,j} + \gamma_{i,j} + \delta_{i,j} + \varepsilon_{i,j} = 1 \tag{4}$$

Furthermore, we define (5)

$$\alpha_{ij} = \frac{e^{\lambda_{a_{ij}}}}{e^{\lambda_{a_{ij}}} + e^{\lambda_{\beta_{ij}}} + e^{\lambda_{\gamma_{ij}}} + e^{\lambda_{\delta_{ij}}} + e^{\lambda_{\varepsilon_{ij}}}} \tag{5}$$

where α_{ij}, β_{ij}, γ_{ij}, δ_{ij} and ε_{ij} are defined by using the softmax function and they are used to control parameters. We use the 1×1 convolution layer to compute the multi-scale coarse depth map's weight, and these weights can learn through standard back-propagation. Experiments have proved that the adaptive module can make the depth map include more detailed structural information.

D. LOSS FUNCTION

In our paper, we use the multiple loss functions proposed by Hu *et al.* [3] rather than the single loss function. The total loss L consists of three parts l_{depth} l_{grad} and l_{normal} shows as (6)

$$L = l_{depth} + l_{grad} + l_{normal}. \tag{6}$$

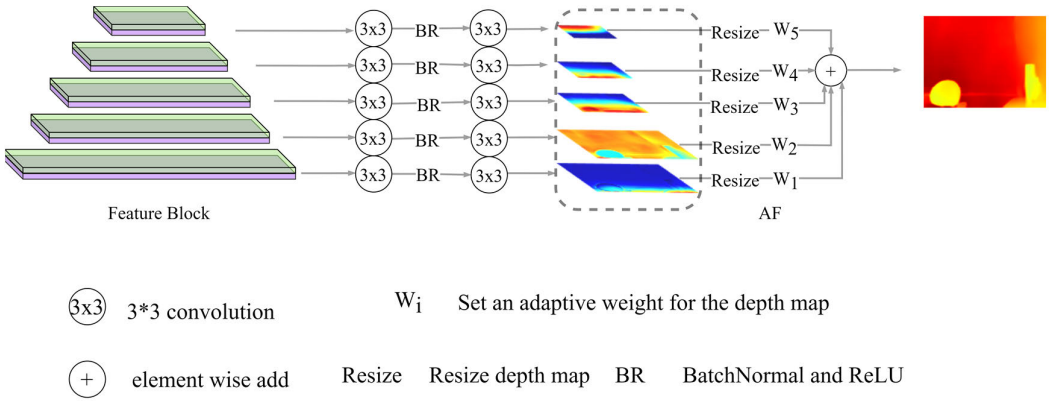


FIGURE 4. The architecture of our DAFM(depth adaptive fusion module). We perform channel compression on each feature block through two 3*3 convolutions to obtain the coarse depth map, and then give each depth map a learnable weight, and then sum the weighted depth maps to obtain a fine depth map.

IV. EXPERIMENT

In this part, we will first introduce the evaluation indicators of the experiment. Then we will introduce the datasets used in the experiment, and conduct various experiments on the datasets to prove the effectiveness of the model.

A. EVALUATION QUANTITATIVE

We use the six metrics propose in prior work [2] to evaluate our model's performance. The six error metrics are defined as

- Root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{y\in T}\left\|\hat{g}_i - g_i\right\|^2} \tag{7}$$

- Absolute relative difference (AbsRel):

$$Abs\,\mathrm{Re}l = \frac{1}{|T|}\sum_{y\in T}\left|\hat{g}_i - g_i\right|/g_i \tag{8}$$

- log10:

$$\log 10 = \frac{1}{|T|}\sum_{y\in T}\left|\log_{10}\hat{g}_i - \log_{10} g_i\right| \tag{9}$$

- Threshold (δ):

$$\%\ \text{of}\ y_i\ \text{s.t.}\ \max(\frac{\hat{g}_i}{g_i}, \frac{g_i}{\hat{g}_i}) = \delta < thr \tag{10}$$

where $thr = 1.25, 1.25^2, 1.25^3$ the g_i is ground truth, $\hat{g}_i$ predict depth value, and T is the available pixels in the ground truth.

B. DATASET AND EXPERIMENTAL SETTING

We mainly conduct experiments on NYU Depth V2 [42]. The NYU-Depth V2 contains a variety of indoor scenes that are most widely used for depth estimation and semantic segmentation. This dataset has 654 aligned RGB-Depth pairs supported to evaluate the model of depth estimation for indoor scenes captured with Microsoft Kinect. In our experiment, we use the training dataset that contains 50K RGB-D images and was preprocessed by Hu *et al.* [3].

We use the Pytorch to implement our model, and then in the encoder state, we use the SENet-154 [31] as our backbone that initialized pre-trained by ImageNet [43].

We set the initial lr $= 0.0003$ and use the learning rate decay policy with Adam optimizer, which will reduce to 10% every 5 epochs, and we set the $\beta_1 = 0.9$, $\beta_2 = 0.999$, epochs $= 10$, and weight decay as 10^{-5}. Our model will be trained on 32GB Tesla V 100 with a patch size is 16.

C. PERFORMANCE COMPARISON

In this section, we evaluate the model that we propose from both qualitative and quantitative points of view on the NYU-Depth V2 dataset. The result has shown that our model achieves state-of-the-art performance.

In TABLE 1, we prepare our model with other previous models to achieve that state-of-the-art evaluation on the NYU-Depth V2 dataset. The result shows that our model obtains the second performance on δ_1, δ_2, and rms the other metrics result shows that our model achieves gains competitive approaches. Qualitative results are illustrated in FIGURE 5. We compare the samples between our model and the previous state-of-the-art model on the NYU-Depth dataset. It is observed that [2] and [7] only have the vague contour of an object. For example, chairs in the (b) [2] and [7] only predict blurry contour but loss most of the structural detail. Although [3] can accurately predict the contour of the chair, it lost some structural detail. For example, the sofa in (d) [3] predicts the contour of the sofa but can't predict the concave-convex of the sofa. Our method not only can predict the boundary of the object clearly but also has shown good performer in the detailed structure. Looking at (c) (d) (f) the bed sofa, and table lamp, our method predicted results have rich detail information and clear boundary than other methods.

D. EDGE ACCURACY COMPARISON

In order to prove that our method can predict detailed information about objects more effectively, we follow Hu *et al.* [3] and use the Precision, Recall, and F1 scores to evaluate the method performance. The result is shown in TABLE 2, we can

TABLE 1. Evaluation results of depth estimation on the NYU V2 test set. The best results are boldfaced, and the second-best ones are underlined. The shown values of the evaluated methods are those reported by the authors in their paper.

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$rel \downarrow$	$rms \downarrow$	$\log10 \downarrow$
Eigen *et al.* [2]	0.611	0.887	0.971	0.215	0.907	-
Liu *et al.* [33]	0.614	0.883	0.971	0.230	0.824	0.095
Cao et al.[34]	0.646	0.892	0.968	0.232	0.819	0.063
Li *et al.* [35]	0.788	0.958	0.991	0.143	0.635	0.063
Laina *et al.* [7]	0.811	0.953	0.988	0.127	0.573	0.055
Xu *et al.* [36]	0.811	0.956	0.987	0.121	0.586	0.052
Ma *et al.*[37]	0.810	0959	0.989	0.143	-	-
Lee *et al.*[38]	0.815	0.963	0.991	0.139	0.572	-
Fu *et al.* [39]	0.828	0.965	0.992	**0.115**	**0.509**	0.051
Qi *et al.* [40]	0.834	0.960	0.990	0.128	0.569	0.057
Hu *et al.*[3]	**0.866**	**0.975**	**0.993**	**0.115**	0.530	**0.050**
Zhang *et al.*[41]	0.833	0.966	0.991	0.132	0.572	0.057
Our baseline	0.833	0.966	0.992	0.131	0.565	0.056
Ours	0.864	0.972	**0.993**	**0.115**	0.525	**0.050**

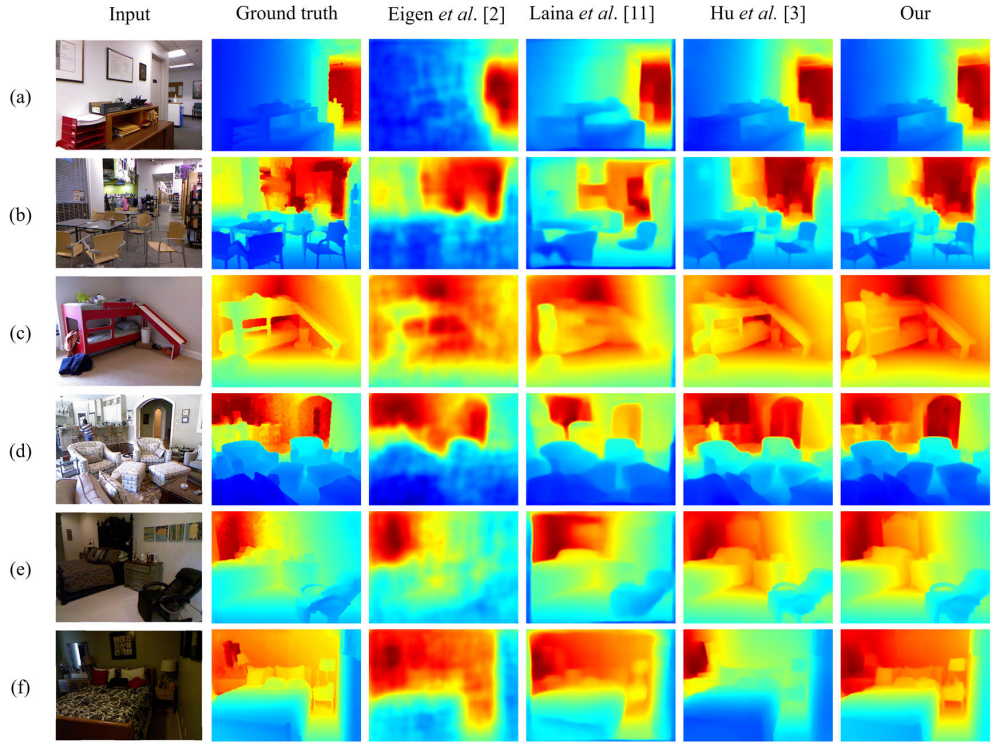


FIGURE 5. Qualitative results on the NYU-D V2 test set. From left to right: input RGB images, ground-truth depth maps, results of Eigen *et al.* [2], Laina *et al.* [7], Hu *et al.* [3], and our method, respectively.

see that our F1 score surpasses all other methods under three different thresholds. As can be seen from TABLE 2. Furthermore, the result shows that our model obtains the second on Recall when the threshold is 0.25 and Precision when the threshold is 1 what's more results show that our model surpasses almost all the existing state-of-the-art approaches in other metrics, This indicates our predictions are more close to the ground truth and have more clear accurate object edges.

E. MODEL TEST IN OTHER DATASET

To further explore the generalization performance of our proposed network, we test the model in the SUN RGB-D dataset [44], which model has trained in NYU Depth V2 [42]. SUN RGB-D contains RGB-D images from NYU Depth V2, Berkeley B3DO [45], and SUN3D [46]. The result is shown in FIGURE 6.

Even though the data distribution of SUN RGB-D and NYU Depth v2 is quite different, our method can still show a good performance. Observed from the (a) group, our estimated not only have clearer edges of the cabinet than other methods predicted but also have a more complete structure than the ground truth. In the (b) and (c) group, our method shows great ability in saving the complete overall scene information. Especially in the (b) group, there is a significant

TABLE 2. Accuracy of recovered edge pixels in depth maps under different thresholds. The best results are boldfaced, and the second-best ones are underlined.

Thres	Method	Prec	Recall	F1
0.25	Laina *et al* [7]	0.489	0.435	0.454
	Xu *et al.* [16]	0.516	0.400	0.436
	Fu *et al.* [39]	0.320	**0.583**	0.402
	Hu *et al.* [3]	0.644	0.508	0.562
	Our	**0.652**	0.518	**0.570**
0.5	Laina *et al* .[7]	0.536	0.422	0.463
	Xu *et al.* [16]	0.600	0.366	0.439
	Fu *et al.* [39]	0.316	0.473	0.412
	Hu *et al.* [3]	0.668	0.505	0.568
	Our	**0.685**	**0.510**	**0.576**
1	Laina *et al.*[7]	0.670	0.479	0.548
	Xu *et al.* [16]	**0.794**	0.407	0.525
	Fu *et al.* [39]	0.483	0.512	0.485
	Hu *et al.* [3]	0.759	0.540	0.623
	Our	0.774	**0.544**	**0.631**

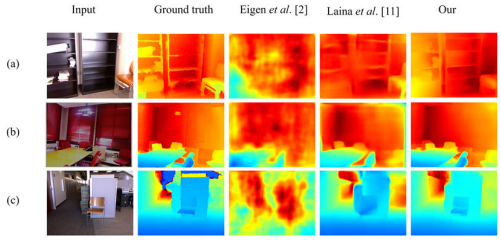


FIGURE 6. The result that the trained model test in SUN RGB-D dataset. From left to right: input RGB images, ground-truth depth maps, results of Eigen *et al.* [2], Laina *et al.* [7], and our method, respectively.

change in the depth value of the corner in our estimation. The results prove that our model has good generalization ability.

F. GENERATE POINT CLOUD FROM DEPTH MAP

To in-depth discuss our work, we design an experiment that we project the estimated depth maps as 3D point clouds. The result has shown in FIGURE 7. The cloud project from our depth map is the closest to the ground truth and our method saves relatively complete background information and performs particularly well in large areas. These point clouds from other methods have more lacks and discontinuities.

G. TIME COMPLEXITY

In this paper, to solve the information loss in the encoder stage, we build an FFP. And to ensure feature information is

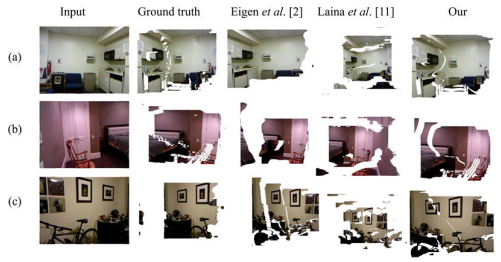


FIGURE 7. The result projects the depth maps as 3D point clouds. From left to right: input RGB images, 3D point projects from ground-truth depth maps, results of Eigen *et al.* [2], Laina *et al.* [7], and our method, respectively.

merged as much as possible, we use the up-sample strategy proposed in [7], instead of pool upsampling, which significantly improves the accuracy of the model. Furthermore, to allow the network to contain more object information, we propose a fusion strategy and channel attention module. The detail is in Sec III. These operations significantly improve the accuracy of the model and enable the estimated depth map to include more object and scene information. However, they also increase the time complexity of the network. Reduce the time complexity and create a lighter model are the goals in our future work.

H. ABLATION STUDIES

In order to further analyze the dense feature fusion network and the depth adaptive fusion module, we conduct two sets of ablation experiments on the NYU Depth V2 dataset. We set the threshold to be 0.5. The result is shown in TABLE 3.

1) DENSE FEATURE FUSION NETWORK

In this part, we compare two sets of experiments with and without DFFN. In our original network, we compressed the original feature map to half of the original amount and then spliced it with the fused feature map to obtain the feature block. To more effectively prove the effect of DFFN, in the network without DFFN, we do not compress the original feature map to ensure that the feature blocks participate in the estimation of the coarse depth map in the two networks that have the same number of channels. As shown in TABLE 3, the result shows that DFFN significantly improves performance. Depth maps were obtained from the two sets of experiments shown in FIGURE 8. In the fifth-scale depth map, the depth map with DFFN has a piece of more precise depth information than without.

TABLE 3. Performance comparison with our module or not. The best results are boldfaced, and the second-best ones are underlined.

Method	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	$rel \downarrow$	$rms \downarrow$	$\log 10 \downarrow$	Prec	Recall	F1
Baseline	0.833	0.966	0.992	0.131	0.565	0.056	0.683	0.460	0.540
Baseline+DAFM	0.862	0.971	0.991	0.118	0.537	**0.050**	0.640	0.442	0.509
Baseline+DFFN	0.856	**0.973**	0.992	0.120	0.535	0.051	0.683	0.499	0.568
Baseline+DAFM+DFFN	**0.870**	0.972	**0.993**	**0.115**	**0.525**	**0.050**	**0.685**	**0.510**	**0.576**

2) DEPTH ADAPTIVE FUSION MODULE

In this part, we introduce the dense feature fusion network (DFFN) like our complete network. In the decoder module, we were decoding step by step to get the depth map. The results are shown in TABLE 3. The result shows that DAFM significantly improves the depth accuracy. The reason for the low edge accuracy is that the depth maps of different scales, directly estimated from the original feature maps, have large gaps in features, resulting in poor results. For example, the first-scale depth map contains almost no depth information, and the fifth scale depth map has almost no object information as shown in FIGURE 8. We connect the original feature with the fused feature map before decoding that to ensure that the feature block of each scale has information of all scales, thereby shortening the feature gap between the depth maps estimated at different scales, and improving the fusion effect. The experiment proves that DFFN + DAFM can achieve huge performance improvements.

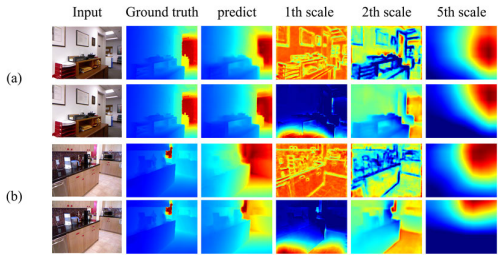


FIGURE 8. (a), (b) are the comparison results of the two groups with DFFN and without DFFN. The first row of each group is without DFFN, and the second row is with DFFN. Each row of pictures from left to right is the original picture, ground truth predicted depth map, First-scale depth map, second-scale depth map, and fifth-scale depth map. First-scale and second-scale depth maps include rich object information. The fifth-scale depth map includes depth information.

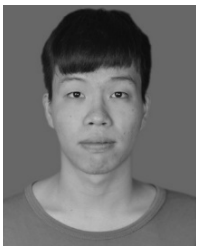
V. CONCLUSION

To deal with the predicted depth lost object information to obtain the depth map that includes rich object information, a Dense feature fusion network, and a novel method to improve the feature fusion effect and reduce the decoder parameters are proposed in our paper. Our main idea is (1) using a dense feature fusion network to aggregate the feature from the encoder to deal with the information lost, (2) designing a new decoder, which can decode multiple feature maps at the same time to obtain multi-scale depth maps, and use adaptive fusion methods to fuse these depth maps to predict fine depth maps that include rich object information. Extensive experimental results demonstrate that the depth maps predicted by our model with more object information than other prework and competitive depth accuracy in the NYU-Depth V2 dataset. However, there are still some problems with our model. First, the time complexity is high. Second, our model is based on supervision and requires a large amount of labeled data. For future work, we will explore a lighter unsupervised depth estimation based on our module.

REFERENCES

[1] C. Tang, C. Hou, and Z. Song, "Depth recovery and refinement from a single image using defocus cues," *J. Modern Opt.*, vol. 62, no. 6, pp. 441–448, Mar. 2015.
[2] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," 2014, *arXiv:1406.2283*. [Online]. Available: http://arxiv.org/abs/1406.2283
[3] J. Hu, M. Ozay, Y. Zhang, and T. Okatani, "Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2019, pp. 1043–1051.
[4] B. Artacho, N. Pandey, and A. Savakis, "Efficient multilevel architecture for depth estimation from a single image," *Electron. Imag.*, vol. 2020, no. 14, pp. 377-1–377-7, 2020.
[5] V. Kaushik and B. Lall, "Deep feature fusion for self-supervised monocular depth prediction," 2020, *arXiv:2005.07922*. [Online]. Available: http://arxiv.org/abs/2005.07922
[6] L. Lin, G. Huang, Y. Chen, L. Zhang, and B. He, "Efficient and high-quality monocular depth estimation via gated multi-scale network," *IEEE Access*, vol. 8, pp. 7709–7718, 2020.
[7] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 239–248.
[8] S. Zhao, L. Zhang, Y. Shen, S. Zhao, and H. Zhang, "Super-resolution for monocular depth estimation with multi-scale sub-pixel convolutions and a smoothness constraint," *IEEE Access*, vol. 7, pp. 16323–16335, 2019.
[9] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
[10] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2650–2658.
[11] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 10, pp. 2024–2039, Oct. 2016.
[12] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, pp. 740–756, pp. 2016.
[13] C. Godard, O. M. Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 270–279.
[14] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*. [Online]. Available: http://arxiv.org/abs/1812.11941
[15] Y. Kim, H. Jung, D. Min, and K. Sohn, "Deep monocular depth estimation via integration of global and local predictions," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 4131–4144, Aug. 2018.
[16] D. Xu, W. Ouyang, X. Wang, and N. Sebe, "PAD-Net: Multi-tasks guided prediction-and-distillation network for simultaneous depth estimation and scene parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 675–684.
[17] Z. Zhang, C. Xu, J. Yang, J. Gao, and Z. Cui, "Progressive hard-mining network for monocular depth estimation," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3691–3702, Aug. 2018.
[18] C. Godard, O. M. Aodha, M. Firman, and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3828–3838.
[19] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, "FastDepth: Fast monocular depth estimation on embedded systems," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 6101–6108.
[20] J. Zhu, Y. Shi, M. Ren, Y. Fang, K.-C. Lien, and J. Gu, "Structure-attentioned memory network for monocular depth estimation," 2019, *arXiv:1909.04594*. [Online]. Available: http://arxiv.org/abs/1909.04594
[21] S. F. Bhat, I. Alhashim, and P. Wonka, "AdaBins: Depth estimation using adaptive bins," 2020, *arXiv:2011.14141*. [Online]. Available: http://arxiv.org/abs/2011.14141
[22] W.-Y. Lo, C.-T. Chiu, and J.-Y. Luo, "Depth estimation from single image through Multi-Path-Multi-Rate diverse feature extractor," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1613–1617.
[23] K. Wu, S. Zhang, and Z. Xie, "Monocular depth prediction with residual DenseASPP network," *IEEE Access*, vol. 8, pp. 129899–129910, 2020.

[24] Y. Chen, H. Zhao, Z. Hu, and J. Peng, "Attention-based context aggregation network for monocular depth estimation," *Int. J. Mach. Learn. Cybern.*, pp. 1–14, Jan. 2021.
[25] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1851–1858.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.
[27] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[28] W. Liu, A. Rabinovich, and A. C. Berg, "ParseNet: Looking wider to see better," 2015, *arXiv:1506.04579*. [Online]. Available: http://arxiv.org/abs/1506.04579
[29] W. Zhou, J. Wu, J. Lei, J.-N. Hwang, and L. Yu, "Salient object detection in stereoscopic 3D images using a deep convolutional residual autoencoder," *IEEE Transactions on Multimedia*, early access, Sep. 21, 2020, doi: 10.1109/TMM.2020.3025166.
[30] J. Wu, W. Zhou, T. Luo, L. Yu, and J. Lei, "Multiscale multilevel context and multimodal fusion for RGB-D salient object detection," *Signal Process.*, vol. 178, Jan. 2021, Art. no. 107766.
[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[32] S. Liu, D. Huang, and Y. Wang, "Learning spatial fusion for single-shot object detection," 2019, *arXiv:1911.09516*. [Online]. Available: http://arxiv.org/abs/1911.09516
[33] F. Liu, C. Shen, and G. Lin, "Deep convolutional neural fields for depth estimation from a single image," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5162–5170.
[34] Y. Cao, Z. Wu, and C. Shen, "Estimating depth from monocular images as classification using deep fully convolutional residual networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 11, pp. 3174–3182, Nov. 2017.
[35] J. Li, R. Klein, and A. Yao, "A two-streamed network for estimating fine-scaled depth maps from single RGB images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3372–3380.
[36] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5354–5362.
[37] F. Ma and S. Karaman, "Sparse-to-dense: Depth prediction from sparse depth samples and a single image," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4796–4803.
[38] J.-H. Lee, M. Heo, K.-R. Kim, and C.-S. Kim, "Single-image depth estimation based on Fourier domain analysis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 330–339.
[39] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.
[40] X. Qi, R. Liao, Z. Liu, R. Urtasun, and J. Jia, "GeoNet: Geometric neural network for joint depth and surface normal estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 283–291.
[41] H. Zhang, Y. Li, Y. Cao, Y. Liu, C. Shen, and Y. Yan, "Exploiting temporal consistency for real-time video depth estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1725–1734.
[42] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, Berlin, Germany: Springer, 2012, pp. 746–760.
[43] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[44] S. Song, S. P. Lichtenberg, and J. Xiao, "SUN RGB-D: A RGB-D scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 567–576.
[45] A. Janoch, S. Karayev, Y. Jia, J. T. Barron, M. Fritz, K. Saenko, and T. Darrell, "A category-level 3D object dataset: Putting the kinect to work," in *Consumer Depth Cameras for Computer Vision*. London, U.K.: Springer, 2013, pp. 141–165.
[46] J. Xiao, A. Owens, and A. Torralba, "SUN3D: A database of big spaces reconstructed using SfM and object labels," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1625–1632.



XIN YANG received the B.S. degree in network engineering from the Guangdong Petrochemical College, in 2019. He is currently pursuing the master's degree in electronics and communication engineering with Wuyi University. His research interests include depth estimation and deep learning.

QINGLING CHANG received the Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in 2015. She is currently a Master's Supervisor and an Associate Professor with Wuyi University. She is also the Sub Decanal of the China-German Artificial Intelligence Institute. She has published more than ten articles included in SCI/EI. In this article, she is mainly responsible for the overall framework design. Her research interests include artificial intelligence, computer vision, and knowledge graph.

XINGLIN LIU received the Master of Computer Technology degree from the School of Computer, Chongqing University, in December 2005, and the Ph.D. degree in applied computer technology from the School of Computer Science and Engineering, South China University of Technology, in June 2012. He is currently an Associate Professor with the School of Innovation and Entrepreneurship, Wuyi University. His current research interests include intelligence computing, text knowledge acquisition, big data, and intelligent recommendation.

SIYUAN HE received the bachelor's degree from Hebei GEO University, in 2018. He is currently pursuing the master's degree in pattern recognition and intelligent systems with Wuyi University. His research interests include 3D computer vision, specifically neural implicit representation for 3D shape reconstruction and 3D scene understanding.

YAN CUI is currently a Professor with the Faculty of Computer Science, Wuyi University. He is also the Dean of the Faculty of Intelligent Manufacturing, Wuyi University. He is also the Dean of the China-Germany Artificial Intelligence Institute. His research interests include computer vision and computer graphic.