IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 30, 2021, accepted August 9, 2021, date of publication August 13, 2021, date of current version August 23, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3104605

Mixed-Scale Unet Based on Dense Atrous Pyramid for Monocular Depth Estimation

YIFAN YANG[1,2], YUQING WANG[1], CHENHAO ZHU[1,2], MING ZHU[1], HAIJIANG SUN[1], AND TIANZE YAN[1,2]

[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]School of Optoelectronics, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding author: Yuqing Wang (wyq7903@163.com)

ABSTRACT Monocular depth estimation is an undirected problem, so constructing a network to predict better image depth information is an important research topic. This paper proposes a mixed-scale Unet network (MAPUnet) with a dense atrous pyramid based on the coder-decoder structure widely used in computer vision. We innovatively introduce the Unet++ structure of the image segmentation network for depth estimation. We reset the number of convolutional layers of the network under the framework of the Unet++ network and innovatively connect the decoders densely. Moreover, by choosing the appropriate size of the atrous radius, we form a dense atrous pyramid based on different feature layers to better connect the features in the deep and shallow layers of the network. To verify the effectiveness of the proposed algorithm, we test the network on the KITTI dataset and the NYU Depth V2 dataset. We compare the network with the current state-of-the-art methods. The proposed method has higher accuracy and has steadily improved relative to the threshold of accuracy and root-mean-square error. We also conduct ablation studies, studies targeting the effectiveness of the network framework, and discussions on the convergence time and parameter complexity of the network. We will open-source the code at https://github.com/yang-yi-fan/MAPUnet.

INDEX TERMS Atrous convolution, dense connection, local and global, multi-scale, pyramid, Unet.

I. INTRODUCTION

The depth prediction method uses image data from a single viewpoint to directly predict the depth value corresponding to each pixel in the image [1]. Depth prediction can be applied to several fields, including robot navigation, autonomous vehicles, or deep space exploration, among other directions, and has a significant impact on 3D imaging technology. Some current sensors can directly detect depth information, such as RGB-D cameras, millimeter-wave radar, LiDAR (Light Detection and Ranging), and ultrasound sensors [2], [3]. Besides, specific depth sensors (e.g., LiDAR sensors) can produce accurate depth measurements at high frequencies. However, due to hardware limitations (such as the number of scanning channels), the depth pixels acquired by these sensors are usually very sparse, which affects their daily use, and they are also more expensive. Therefore, there is still a gap to applicate pervasively of Lidar to the task of a scene of 3D information.

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Wei.

The depth estimation problem is ill-posed; in other words, the image depth solution is not unique, e.g., we can recover many 3D scenes from a 2D RGB image [4]. Thus, to understand the scene 3D geometry from a single image, one considers not only local cues such as texture appearance information under various lighting occlusion conditions, viewpoint information, or scale information relative to known objects to obtain the geometric object parameters but also global contextual cues based on a statistical perspective to obtain scene information such as the overall shape or layout of the scene [5]–[8].

Classical computer vision approaches use multi-view stereo correspondence algorithms for depth estimation [9]. With the rapid development of deep learning in the last decade, considerable progress has been made in research for deep estimation tasks [10]–[13]. In deep learning, the monocular depth estimation problem can be described as a dense pixel-level continuous regression problem or modeled as a classification [14] or quantile regression [1]. Although current semi-supervised [15] or unsupervised learning methods [1], which do not rely exclusively on ground truth depth data, have made some progress. However, semi-supervised

or unsupervised depth estimation is still not as effective as deep convolutional neural network (DCNN) models with supervised approaches.

Inspired by the structure of Unet++ and Unet3+, we propose a supervised network incorporating a dense atrous pyramid structure, which we call MAPUnet, as shown in Fig. 1(a). Specifically, this structure uses ResNet-101/Densenet-161 as an encoder. It accesses convolutional layers of 1, 2, and 3 node numbers at each encoding stage with spatial resolutions of 1/8, 1/4, and 1/2, respectively, to form a shallow-to-depth encoder-decoder transducer structure which we refer to as a transducer in this paper. Based on this, we densely connect the decoder structure from deep to shallow layers. Then, inspired by DenseASPP [16], we replace the convolutional layer in the transducer with the atrous convolutional layer. To improve the information flow throughout the network and facilitate better gradient transfer, we add implicit depth supervision modules between the nodes at the shallowest level of the transducer and between the decoder nodes from deep to shallow. The innovation of this paper is to connect the coding stages of different resolutions of the underlying network (ResNet or DenseNet) to the corresponding dense atrous pyramid layer and integrate the information utilizing a dense connection decoder. We build a multi-scale fusion and feature pyramid structure through a deep enough network with a layer-by-layer codec structure. The structure helps the network develop information at different scales and connect features at various levels with different resolutions nonlinear function relationship between images and depth effects. We put the MAPUnet network to experiments on the KITTI dataset [17] and NYU Depth V2 dataset [18]. The experiments show that the method reaches a more advanced level.

Contributions: Our main contributions are the following:

- To the best of our knowledge, we introduce for the first time the Unet++ segmentation network structure to the monocular depth estimation work—a comprehensive integration of global information with local information.
- We further densely connect the decoders so that both the transitional part and the decoder part of the network are densely connected to achieve implicit deep supervision of the decoder part.
- We replace the middle transducer part with a dense atrous pyramid structure. Through the superposition of convolutional layers with different atrous radii, the pyramid structure can fuse large-scale information with small-scale information, allowing the network to estimate the depth contours of objects at different scales.

The rest of this paper is organized as follows. In Section II, we present the related work, and in Chapter III, we describe the proposed approach in detail. In Section IV, we conduct algorithm comparison experiments, ablation experiments, etc. We analyze the effect of the core factors of the proposed method on network effectiveness. We perform quantitative and qualitative analyses. In Section V, we discuss the shortcomings of the proposed method. Finally, in Section VI, we conclude the article.

II. RELATED WORK

A. SUPERVISED MONOCULAR DEPTH ESTIMATION

Earlier work on depth estimation mostly estimated depth information by studying the point correspondence between images and triangulation [9]. For example, Saxena *et al.* [7] used Markov random fields (MRF) to extract absolute depth magnitudes and relative depth magnitudes between objects in the scene using local cues at multiple scales, making full use of local and global information of the image blocks. Since hand-crafted features alone can only capture local information, probabilistic graphical models, such as MRFs, are usually built on top of these features to incorporate local cues from long distances to form global cues. Saxena *et al.* [6] also performed depth estimation based on the assumption that the 3D scene contains many small planes (i.e., triangulation) by estimating the 3D position and orientation of the over-segmented superpixels in the image. Later work, such as the DepthTransfer method [8], had successfully used global information to find candidate images using GIST global scene features, where the candidate images were very similar to the input RGBD images in the database. Since then, various methods had been proposed for depth estimation using hand-crafted cues [6], [19]–[21] while also incorporating long-range and global cues [22]. However, manually labeled cues have limitations and do not cope well with changes, including rotation, scaling, etc. With the continuous iterative update of the technology, many current depth estimation benchmarks are based on neural network algorithms [1], [23].

Eigen *et al.* first proposed the use of deep learning to solve the monocular depth estimation problem in [10], describing how to train a network on sparse labels obtained from LiDAR scans to estimate depth from a single image. Rapidly, Eigen *et al.* further predicted by a global coarse depth network coupled with a fine-segmentation network targeting local regions, and unlike previous work on single image depth estimation, this network could learn representations from the original image pixels without some hand-crafted features such as contour lines, superpixels, or low-level segmentation [14]. These ideas were gradually developed into architecture [13] and training techniques [12], [15]. Li *et al.* [24] predicted superpixel depth maps by Convolutional Neural Network (CNN) models and then used Conditional Random Field (CRF) to refine the depth maps to the pixel level. DCNF network proposed a unified approach combining CRF and fully convolutional networks based on superpixel pooling methods to speed up inference [12], [25] used the multi-scale outputs of the different phases of the CNN to fuse with the continuous CRF outputs in [26] to further refine the depth information through the attention module on the feature map. However, there is already information loss by converting RGB images into superpixel images, and more information may be lost in the features extracted after pooling them. Although the method using CRF (conditional random field) can improve the depth prediction ability of the network, it leads to a massive dimensionality of the input, and it can be tough to construct the probability distribution among

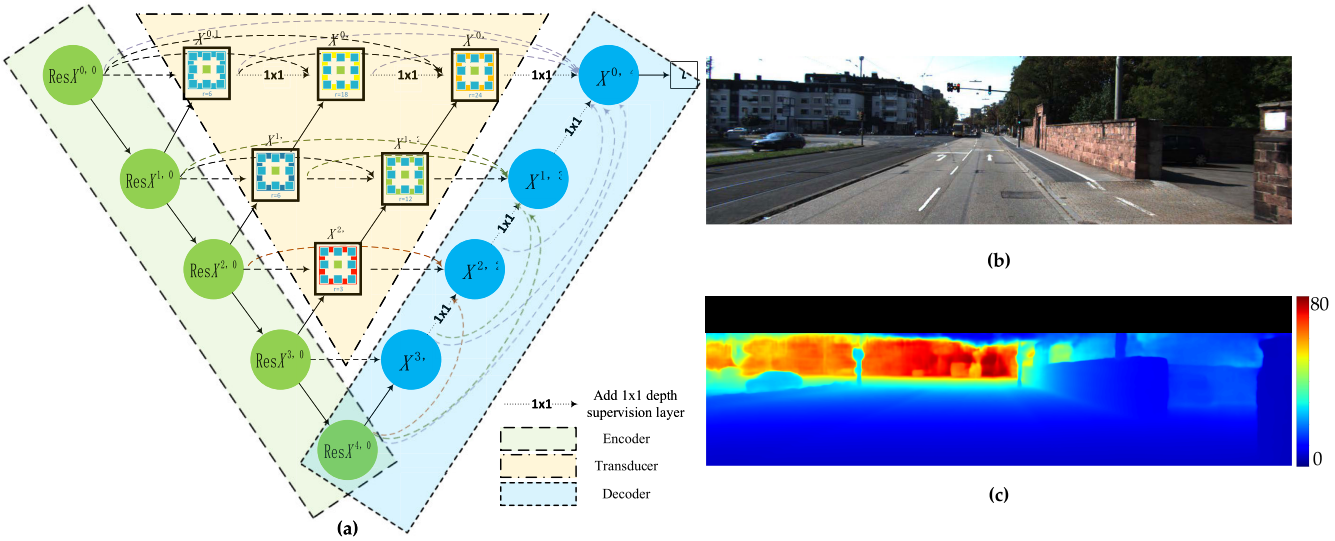

FIGURE 1. Figure (a) shows an overview of the proposed network MAPUnet's architecture, where the units of the numbers in the pseudo-color scale bar are meters (m). We refer to the node operation between the encoder and the decoder as a transducer. Hence, the network consists of a feature encoder, a transducer (a dense atrous convolutional pyramid layer), and a densely connected decoder. We use skip connections to achieve the flow of information in the network. Figure (b) shows the original RGB map of the input network, and Figure (c) represents the depth map of the network prediction.

the features due to the complex dependencies between the features.

More and more scholars have tried to use encoder-decoder network structures to predict depth values in recent years. DenseDepth [27] used a pre-trained DenseNet [28] as the backbone, with bilinear upsampling and skip concatenation on the decoder to obtain a high-resolution depth map. The novelty of the Bts [4] architecture is that its Local Planar Guidance (LPG) module can replace the upsampling-based skip connection module to convert intermediate feature maps into full-resolution depth predictions. The network achieves better estimation results, but the number of network parameters is large. BANet [29] proposed a lightweight bi-directional attention network that filters ambiguous information from in-depth features by combining global and local contextual information. The network reduces network parameters with minimal loss of estimation accuracy, but the network is more complex. Xia *et al.* [30] proposed a generalized task-independent monocular model that outputs a probability distribution of its scene depth based on the input color image and outputs a sample approximation a VAE (Variational Auto-Encoder). The model increases the dimensionality of the network while achieving better results. Aich [29] *et al.* used a pairwise ranking loss to bootstrap sampling point pairs through low-level edge maps and high-level object instance masks to efficiently learn depth estimates from ground truth depth data. Still, this loss function is less effective when trained under sparser ground truth.

In this paper, we propose a deep neural network-based framework model that avoids manual feature extraction. In the Unet++ framework, the innovative change of the message transfer structure of the decoder and the extraction of features using densely connected atrous convolutional layers achieve better depth estimation. Furthermore, this network has a reduced number of weights compared to the Bts network.

B. SELF-SUPERVISED MONOCULAR DEPTH ESTIMATION

In recent years, self-supervised monocular depth estimation has been proposed. For this problem, Garg *et al.* [32] and Godard *et al.* [33] proposed an ingenious solution that indirectly transforms the direct depth estimation problem into an image reconstruction problem. They converted the depth estimation modeling problem into an issue of the geometric properties of the image projection transform between stereo image pairs. In other words, the network can be optimized according to the photometric error between the projected image and the actual image. Subsequently, Zhou *et al.* [34] showed that the depth information and relative pose information between two video frames could be predicted simultaneously utilizing a joint optimization network. Based on this idea, networks have been optimized from several perspectives, such as the improved loss function proposed by Aleotti *et al.* [35], the targeted network structure proposed by Guizilini *et al.* [36], the innovative approach of mixing video and stereo data by Zhan *et al.* [37], and the improved optimization by Casser *et al.* strategies [38], among others. DORN [1] models the MDE(Monocular Depth Estimation) task as an ordered regression problem with a spatially increasing (logarithmic) discretization of the depth range to appropriately reduce the error that increases with increasing depth values. The algorithm presents novel ideas, but there is still much room for improvement in algorithm effectiveness. Godard *et al.* [39] proposed a very advanced algorithm that uses minimum reprojection loss to deal with occlusion between different frames. However, it is not possible to consider the depth information of the mismatched scene positions between two frames. Eldesokey *et al.* [40] proposed

a probabilistic version of the Normalized Convolutional Neural Network (NCNN), which learns the input confidence estimator utilizing self-supervision to identify the input interference noise. The model is small and achieves better results in the unsupervised domain, but still has a gap compared to the supervised algorithm.

In the past two years, more and more scholars have used adversarial discriminant learning methods to predict depth information with better results. References [41] trained an encoder that extracts imperceptible nighttime features that distinguish daytime images by an adversarial discriminative learning method based on PatchGAN, and plug a pre-trained daytime depth decoder into its back end to achieve unsupervised nighttime monocular depth estimation. S3Net [42] considered the geometric structure across space and time in monocular video frames in an adversarial network framework, i.e., using geometric, temporal, and semantic constraints simultaneously for depth prediction. However, adversarial discriminative learning methods generally require better pseudo labeling to complete training, and this problem still requires supervised methods to solve.

Good results have also been achieved by semantic segmentation to guide depth estimation, and SGDNet [43] proposed a cross-domain training model to guide unsupervised depth estimation by supervised semantic segmentation. Moreover, SGDNet showed by the study that the mask information of semantic segmentation would effectively prevent the contamination of photometric loss from moving objects. However, only part of the results obtained from semantic segmentation and depth estimation overlap, and the part that does not overlap may affect depth estimation results.

The model in this paper has more weights than the unsupervised model and is not as compact as most unsupervised models, which is the area to be improved in this algorithm.

C. ENCODER-AND-DECODER STRUCTURE AND ATROUS CONVOLUTION STRUCTURE

In 2015 Unet won the competition and significantly improved the ISBI cell tracking challenge [44]. The codec structure of its network was the key to performance improvement. Since then, the latest results based on deep neural networks (DCNN) have been typically divided into two parts: an encoder for intensive feature extraction and a decoder for predicting the task outcome [12], [33]. Dense feature extractors usually use very powerfully underlying deep networks such as VGG [45], ResNet [46], or DenseNet [28]. Moreover, the decoders are designed as appropriate depending on the task. In 2018, Unet++ [47] improved on Unet by introducing nesting and dense connections to enhance the transfer of information in the network and reduce the semantic gap between encoders and decoders. Recently, Unet3+ [48] introduced full-scale skip connections to more fully exploit multi-scale features, combining low-level details with high-level semantic information to obtain full-scale feature maps. In 2016, dilated convolutions were first proposed and applied in image segmentation. As the research progressed, atrous convolution pyramid pools were used in semantic segmentation [49] and depth estimation [1], [4]. Since the dilated convolution allows a larger receptive field, sparse convolution with different radii of the atrous rate can capture large-scale variations in the image and improve the network's performance. The encoder-and-decoder structure increases the complexity of the network, and since the network is deeper, the features extracted are more comprehensive, allowing us to train the network better.

In this paper, we borrow the encoder-decoder structure, innovatively introduce the Unet++ and Unet3+ structures into the depth estimation, and use the atrous convolution to help the network refine the features.

III. METHODS

The MAPUnet network proposed in this paper is a deep enough network. The network uses the structure of encoder+transducer+decoder can effectively improve the depth estimation accuracy, and the network mines the information in a single RGB image through different stages of the encoder. Furthermore, the network integrates many valuable features through the transducer and decodes them through the decoder. A single RGB image is fed into the MAPUnet network, and the network outputs a depth map of the corresponding image.

A. OVERALL NETWORK ARCHITECTURE

Fig. 2(a) depicts the network's overall framework, an innovative new network based on the Unet, Unet++, and Unet3+ networks. Unet network is the basis of Unet++ and Unet3+ networks. However, although the Unet decoder part shares the same encoder, the Unet decoder is disconnected, i.e., the contents of the deep U-shaped network encoder nodes do not provide supervisory signals to the decoder corresponding to the shallow stage in the upper layer. Also, fusing the feature maps of the decoders in the network with the encoder same-scale feature maps does not guarantee that the same-scale feature maps are the best match for feature fusion. Therefore, Unet++ was born, as shown in Fig. 2(b). Unet++ removes the original coder-decoder skip connection in Unet and connects all neighboring nodes within the ensemble, which realizes the information linkage of multi-level networks and increases the possibility of achieving optimal feature fusion. Unet3+ is even more innovative in adding feature fusion operations from large-scale to small-scale and connects decoders equally densely.

Although Unet3+ achieves better results on top of medical image segmentation compared to Unet++, in terms of depth estimation, the depth of most regions in the image is progressively changing. Therefore, it is not regional like image segmentation is. Thus, the fusion of more significant size features with smaller size features may not improve the network's prediction of depth values. Moreover, similar to the Unet++ network, the features are progressive layer by layer and change gradually, analogous to the gradual change in depth, which helps the network understand the depth

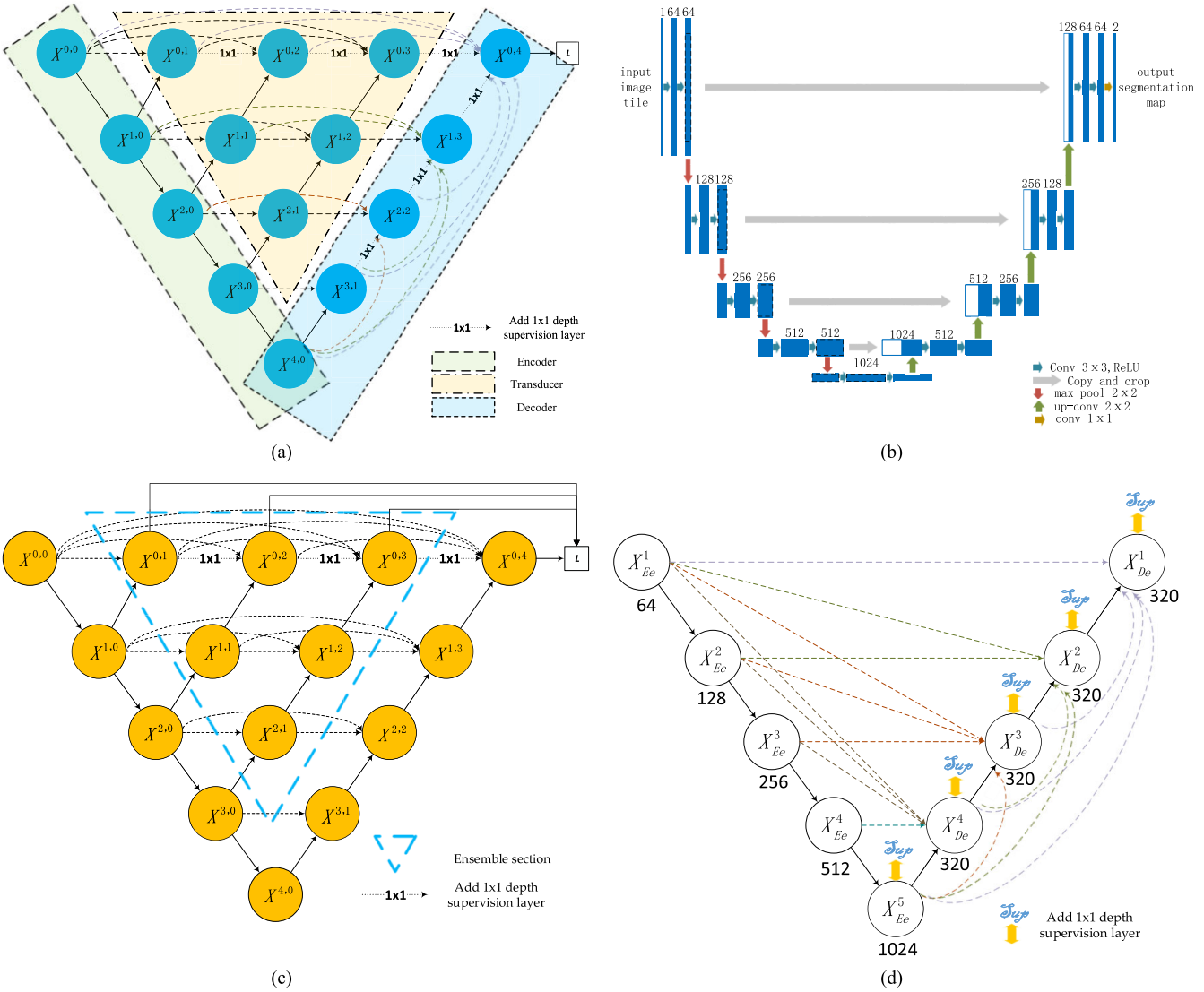

FIGURE 2. Fig. (a) represents the Unet++ network that connects the decoders densely. Fig. (b) illustrates the network structure of Unet. Fig. (c) depicts the network structure of Unet++. Finally, figure (d) represents the network structure of Unet3+, which we call MUnet. In the field of medical segmentation, the optimization of the well-known Unet structure is mainly based on the addition of the multi-scale fusion module and the addition of the dense fusion operation of different nodes at the same scale. In this paper, the network of this paper is based on Unet++ and innovatively introduces the structure in Unet3+ decoder to obtain the MUnet network structure.

information better and make predictions. For comparison experiments, please see the EXPERIMENTS section.

Borrowing from the Unet++ network structure and the Unet3+ network structure, we connect all the layers (feature maps with matching size characteristics). Thus each shallow layer receives additional input from all previous deeper layers and passes its feature map to the more external subsequent network layers. This operation maximizes the flow of information among the layers in the network. At the same time, making each node between the layers skip connected, the traditional feed-forward architecture can be seen as a state algorithm where the state is passed layer by layer. Each layer reads the state from its predecessor layer and writes to the following layer. The network at each level changes the state but also passes information that needs to be saved.

In this paper, we borrow the Unet++ structure and use ResNet or DenseNet as the encoder for feature extraction of the input image, the dense atrous pyramid as the transducer, and a similar dense connection decoder position. While not drawing representation power from extremely deep or wide architectures, this structure can produce highly efficient condensed models about parameters by exploiting the potential of the network through feature reuse. At the same time, concatenating the feature maps learned at different layers increases the variability of the input at subsequent layers and improves the efficiency of the network parameters. In addition to better parameter efficiency, the decoder structure in this paper likewise improves the information flow and gradients across the network, making it easy to train. Each layer has direct access to the gradients from the loss function and the original

input signal, enabling implicit deep supervision, facilitating the training of deeper network architectures. In addition, borrowing from the deep supervised operation in Unet++, we add convolutional layers with a convolutional kernel size of 1×1 at the shallowest transducer node and at the decoder node to avoid the gradient of the network from becoming zero at deeper parts of the network. Besides, the dense connectivity of the network has a regularization effect, reducing the possibility of overfitting phenomena on tasks with smaller size training sets.

Suppose $x^{i,j}$ denotes the output of the $X^{i,j}$ node, where i denotes the sequence number of the down-sampling layer of the encoder and j denotes the sequence number of the nodes in each layer (dense block) from the encoder to the decoder. $x^{i,j}$ represents the node operation of the feature map as

$$x^{i,j} = \begin{cases} H\left(D\left(x^{i-1,j}\right)\right), & j = 0 \\ H\left(\left[D\left(x^{i+1,j-1}\right), \left[x^{i,k}\right]_{k=0}^{j-1}\right]\right), & i+j<4 \cup j \neq 0 \\ H\left[D\left(x^{i+1,j-1}\right), \left[x^{i,k}\right]_{k=0}^{j-1}, \left[U\left(x^{4-l,l}\right)\right]_{l=0}^{j-1}\right], & i+j=4 \cup j \neq 0 \end{cases} \tag{1}$$

where the function $H(\cdot)$ is the operation representing the convolution operation + activation function, $D(\cdot)$ and $U(\cdot)$ denote the down-sampling operation and up-sampling operation, respectively, and $[\cdots]$ suggests the concatenated layer operation. As shown in Fig. 2(a), when $j = 0$ in the node, the node receives only the input from the previous stage of the encoder, and when $j = 1$ in the node, the node receives two inputs from the encoder sub-network on two successive levels, which can be summarized as follows: when $j > 0$ and $i + j < 4$, the node receives the input from $j + 1$ nodes, where j inputs are the outputs of the first j nodes in the same level, and the $(j + 1)$th input is the up-sampled output of a deeper layer of skip connections. When $j > 0$ and $i + j = 4$, the node receives $2j$ inputs, including skip connections in the same hierarchy, multiple skip connections from deeper layers.

In the encoder, we use ResNet-101/DenseNet-161, remove its last two layers: the average pooling layer, and the fully connected layer, and pass five blocks as output nodes to the subsequent nodes.

B. DENSE ATROUS SPATIAL PYRAMID POOLING MODULE

Depth prediction, like image segmentation, belongs to the same category of dense prediction, i.e., depth information is predicted for each pixel location in an image. Therefore, how to better utilize the contextual content information in an image is a problem worth investigating. The concept of atrous convolution was proposed in the paper of [49], which exponentially expands the receptive field with the loss of resolution or coverage through a rectangular prism of convolution layers. The atrous convolution operator was called "convolution with dilation filter" in the past, and it is equivalent to using the filter parameters differently. Although the filter parameters are the same, the size of the location where the dilation is done centered on a particular pixel point is different. Thus the value after convolution with the filter is different. The atrous convolution operator plays a crucial role in the algorithm, a wavelet decomposition [50]. In addition, the atrous convolution layer is capable of multi-scale contextual aggregation.

In the one-dimensional case, let $y[i]$ represent the output signal, and $x[i]$ denotes the input signal. Then, the atrous convolution can be described as follows:

$$y[i] = \sum_{k=1}^{K} x[i + d \cdot k] \cdot w[k] \tag{2}$$

where d denotes the atrous radio, $w[k]$ represents the parameter of the kth filter, and K means the size of the filter. DeepLabV3 [51] proposed ASPP (Atrous Spatial Pyramid Pooling),i.e., parallel atrous convolution mode. The parallel atrous convolution layer consists of multiple atrous layers, each receiving the same input. The outputs of the convolution layers are also cascaded together, as shown in the following equation:

$$y = H_{3,6}(x) + H_{3,12}(x) + H_{3,18}(x) + H_{3,24}(x) \tag{3}$$

[16] proposed the dense atrous convolution, i.e., DASPP module, expressed as Eq:

$$y_l = H_{K,d_l}([y_{l-1}, y_{l-2}, \ldots, y_0]) \tag{4}$$

where d_l represents the expansion rate of layer l, $[\cdots]$ represents the cascade operation, and $[y_{l-1}, \ldots, y_0]$ represents the feature map that connects the outputs of all previous layers. DenseASPP allows more pixels to participate in the computation of the feature pyramid than the normal convolution operation and ASPP operation. By skip-connecting shared atrous convolution layers, convolution layers with large atrous rates and convolution layers with low atrous rates will work interdependently. As a result, the DenseASPP structure will get a denser feature pyramid and a larger receptive field to perceive enormous background information.

This paper proposes a densely connected module with more scales and calls it MLDenseASPP (Multi-Layers DenseASPP), as shown in Fig. 3 below. We densely connect the node outputs within the same layer at multiple scales, but also, between layers, we build skip connections to make the connections between feature layers dense as well. The MLDenseASPP module built in this paper is (3, 6, 6, 12, 18, 24), and the corresponding node positions are $\left(X^{2,1}, X^{1,1}, X^{0,1}, X^{1,2}, X^{0,2}, X^{0,3}\right)$. According to [16], for an atrous convolution layer with atrous rate d and kernel size K, the receptive field size is equivalent to the following equation:

$$R = (d - 1) \times (K - 1) + K \tag{5}$$

Assuming that the dimensions of the two convolutional layers of the stack are K_1 and K_2, respectively, the new

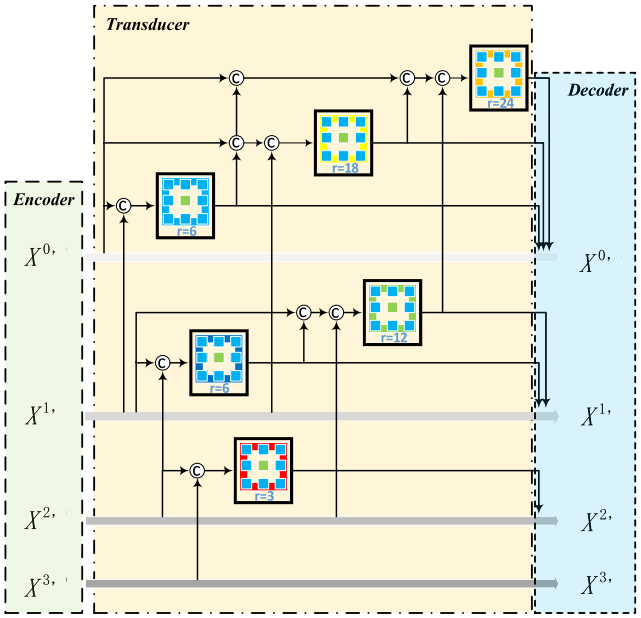

FIGURE 3. In the transducer, the output of each atrous convolution layer is concatenated with the input feature layer and then fed to the next node for processing. The feature maps of different scales are expanded to the corresponding size using nearest-neighbor interpolation. Where © represents the channel concatenation.

receptive field size is

$$K = K_1 + K_2 - 1 \tag{6}$$

Therefore, in the MLDenseASPP (3, 6, 6, 12, 18, 24), the equivalent maximum receptive fields are

$$R_{\max} = R_{3,3} + R_{3,6} + R_{3,12} + R_{3,18} + R_{3,24} - 5 = 139 \tag{7}$$

such a large receptive field enables the extraction of global information about larger objects in the feature image.

C. FUSION OF MORE LAYERS OF FEATURES

In each layer, there are fusion features with different block feature layers with fusion ratios of 1/5, 1/5, 1/3, 7/55, 5/21, and 13/84, respectively. For the generated depth features, if the feature size is small, e.g., size (batch_size, 44, 88), a smaller atrous convolution layer of radius three is used for processing. As the feature layer size becomes larger, we gradually increase the radius in the atrous convolution to obtain better global and background information. As shown in Fig. 4 below, the number in each strip represents the radius value, the length of each strip represents the equivalent kernel size for each combination, and the shaded area represents the proportion of the contribution of deep features to shallow features. Thus, the dense connection between multiple stacked atrous convolution layers can form a feature pyramid with more dense and diverse scales.

For example, consider the case of a dense atrous pyramid with a minimum number of nodes, as shown in Fig. 5(a) below. The contribution ratio of the deep level to the shallow level in a pyramid with only three nodes, $X^{0,1}$, is $P\left(X^{0,1}\right) = 32/(64+32) = 1/3$. And in Fig. 5(b), the contribution ratio of deep level features in $X^{1,1}$ is $P\left(X^{1,1}\right) = 64/(256+64) = 1/5$, then the contribution ratio of deep level to shallow level in $X^{0,2}$ is $P\left(X^{0,2}\right)$:

$$\begin{aligned} P\left(X^{0,2}\right) &= \frac{C^{0,1}P\left(X^{0,1}\right) + C^{1,1}P\left(X^{1,1}\right)}{C^{0,0} + C^{0,1} + C^{1,1}} \\ &= \frac{32 \times 1/3 + 32 \times 1/5}{64 + 32 + 32} = \frac{5}{21} \end{aligned} \tag{8}$$

where $C^{a,b}$ denotes the number of channels output by the (a, b)th node in the network. As the number of nodes in the pyramid increases, more in-depth information will converge into the shallow network. At the same time, expanding the radius of the atrous convolution can help convolutional layers with lower sampling rates to sample multi-layer feature pixels more intensively. As a result, this operation improves the information flow between networks and facilitates the better transfer of scale information.

D. TRAINING LOSS

Both [10] and [4] used the following loss functions:

$$D(g) = \frac{1}{T}\sum_i g_i^2 - \left(\frac{1}{T}\sum_i g_i\right)^2 + (1-\lambda)\left(\frac{1}{T}\sum_i g_i\right)^2 \tag{9}$$

where $D(g)$ can be simplified as

$$D(g) = \frac{1}{T}\sum_i g_i^2 - \frac{\lambda}{T^2}\left(\sum_i g_i\right)^2 \tag{10}$$

where $g_i = \log \tilde{d}_i - \log d_i$, d_i is the ground truth depth, λ is a constant, and T denotes the number of pixels with valid truth values.

According to equation (9), it can be seen that $D(g)$ is the sum of the variance and the weighted squared mean of the logarithmic space of the depth image. Increasing the value of λ allows $D(g)$ to focus more on variance minimization. Finally, we also follow the loss function formula in the final [4], i.e.

$$L = \alpha\sqrt{D(g)} \tag{11}$$

where α is set to improve convergence, such that $\alpha \equiv 10$.

IV. EXPERIMENTS

To verify the effectiveness of this network, we test it on top of a challenging benchmark dataset and show our results. The tests are performed on top of the KITTI dataset and NYU Depth V2 dataset. In addition, we will conduct ablation experiment comparisons and some other experimental comparisons in the subsequent sections.

A. IMPLEMENTATION DETAILS

We use the open-source deep learning framework PyTorch [52] to train our network. For training, we use the Adam optimizer [53] and set the learning rate to 10^{-4}, epoch to 50, and batch size to 2 on a computer with two NVIDIA 2080ti GPUs. We changed the last three layers in the decoder to deformable convolution [54]. Replacing specific layers

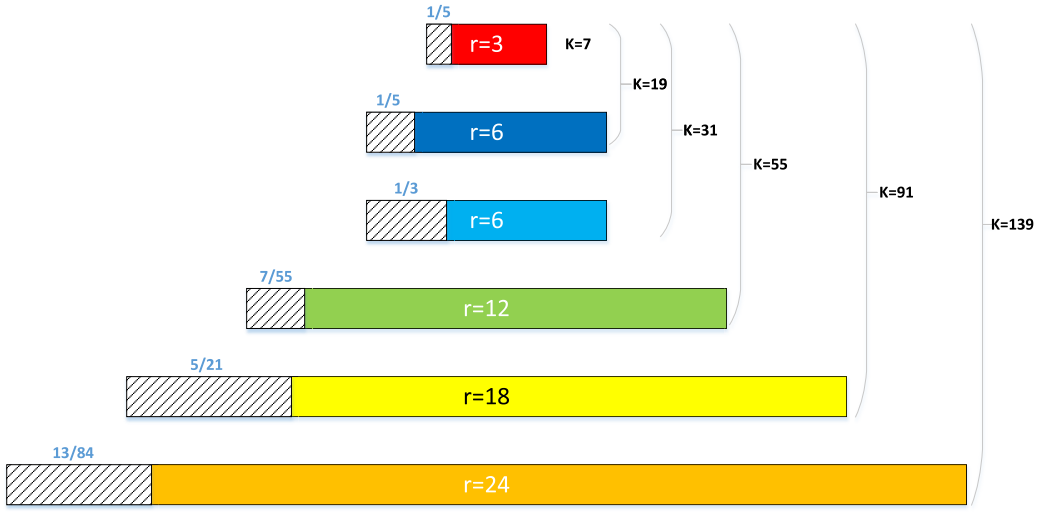

FIGURE 4. Illustration of the scaled pyramid in the transducer. In the dense atrous convolutional layer, the radius of dilation is the deepest radius of the transducer r = 3, respectively, which is connected to the output of the third stage encoder. The next deepest layer is r = 6, 12 from near the encoder to near the decoder direction, respectively, where r = 6 connects to the output of the second encoder. Finally, the shallowest layer is r = 6, 18, 24, where r = 6 connects to the output of the shallowest encoder. The shaded part of the figure indicates the proportion of the contribution of deep features to shallow features. K denotes the radius of convolution of equivalent atrous composed of feature pyramids.

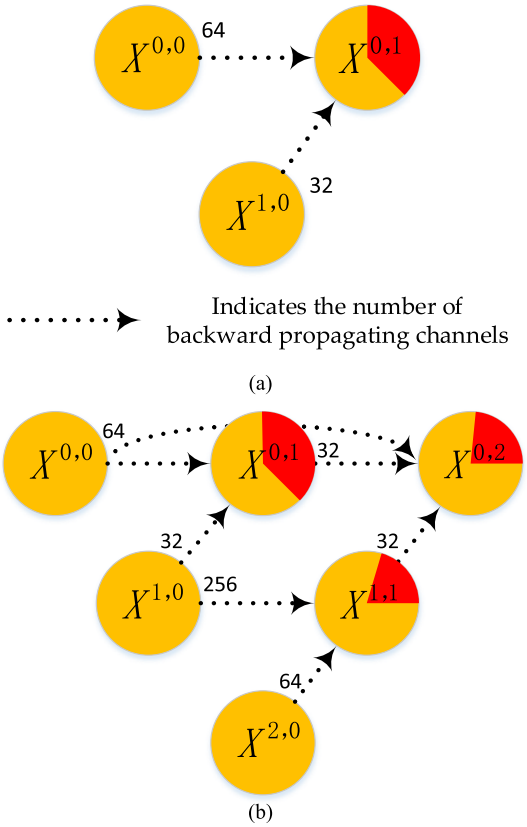

FIGURE 5. The abbreviated pyramid model, where the arrows indicate the number of channels propagating backward. The red area shows the proportion of the contribution of deep features to shallow features.

with deformable convolution in the network improves the network's performance in semantic segmentation and target detection. Therefore, to improve the MAPUnet network performance, we changed the last three layers of the decoder to deformable convolution following the operation in the paper [54]. Deformable convolution adds its offset to each pixel point in the convolution kernel and learns the offset from the target task without additional supervision. Deformable convolution can generalize various transformations such as scale, (anisotropic) aspect ratio, and rotation.

For the backbone selected in the network, either ResNet-101 [46] or DenseNet-161 [28], are pretrained weight models about image classification on the ILSVRC dataset [55]. We use random horizontal flips and random contrast, brightness, and color adjustments to enhance the images in the data preprocessing part. Also, we perform random rotation operations on the images to increase the robustness of the trained network. It takes about 25 hours to train 25 epochs on the KITTI dataset and about 48 hours to train 25 epochs on the NYU Depth V2 dataset.

B. EVALUATION METHOD

For the evaluation, we use the evaluation metrics recommended by the KITTI dataset:

- Threshold(Accuracy):

$$Accuracy = \%\ \text{of}\ \tilde{d}_i\ \text{s.t.}\ \max\left(\frac{\tilde{d}_i}{d_i}, \frac{d_i}{\tilde{d}_i}\right) = \delta < thr \quad (12)$$

three different thresholds (1.25, 1.25^2, 1.25^3) are used to measure the accuracy of the predicted depth values.

- Absolute Relative Error:

$$\text{Abs Rel} = \frac{1}{|T|}\sum_{\tilde{d}\in T}\frac{\left|\tilde{d} - d\right|}{d} \quad (13)$$

TABLE 1. Performance comparison on KITTI Eigen split, set at a distance of 0-80m.

Method	cap	higher is better			lower is better			
		δ<1.25	δ<1.25^2	δ<1.25^3	Abs Rel	Sq Rel	RMSE	RMSE *log*
Make3d [6]	0-80m	0.601	0.820	0.926	0.280	3.012	8.734	0.361
Eigen et al. [10]	0-80m	0.702	0.898	0.967	0.203	1.548	6.307	0.282
Liu et al. [12]	0-80m	0.680	0.898	0.967	0.201	1.584	6.471	0.273
Godard et al. [34]	0-80m	0.861	0.949	0.976	0.114	0.898	4.935	0.206
DORN [1]	0-80m	0.932	0.984	0.994	0.072	0.307	2.727	0.120
Yin et al. [56]	0-80m	0.938	0.990	0.998	0.072	-	3.258	0.117
LPF [57]	0-80m	0.7147	0.8996	-	0.2033	-	6.5613	-
DenseDepth [28]	0-80m	0.886	0.965	0.986	0.093	0.589	2.727	0.120
Bts-Resnet101 [4]	0-80m	0.954	0.992	0.998	0.061	0.261	2.834	0.099
Bts-Densenet161 [4]	0-80m	**0.955**	**0.993**	0.998	**0.060**	0.249	2.798	**0.096**
MAPUnet-Resnet101	0-80m	**0.955**	**0.993**	0.998	0.0620	0.250	**2.708**	0.097
MAPUnet-Densenet161	0-80m	**0.955**	0.992	**0.999**	0.061	**0.242**	2.741	**0.096**

- Square Relative Error:

$$\text{SqRel} = \frac{1}{|T|}\sum_{\tilde{d}\in T}\frac{\left\|\tilde{d}-d\right\|^2}{d} \tag{14}$$

- Root Mean Square Error:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum\left(\tilde{d}-d\right)^2} \tag{15}$$

- Root Mean Square logarithmic Error:

$$\text{RMSE log} = \sqrt{\frac{1}{|T|}\sum_{\tilde{d}\in T}\left\|\log\tilde{d}-\log d\right\|^2} \tag{16}$$

here T refers to the number of pixels with valid truth values.

- Scale Invariant logarithmic Error:

$$SILog = \frac{1}{n}\sum_i^n y_i^2 + \frac{1}{n^2}\left(\sum_i^n y_i\right)^2 \tag{17}$$

where $y = \log\tilde{d} - \log d$.

C. KITTI DATASETS

The KITTI dataset for monocular depth estimation [17] is a subset of the KITTI family of datasets designed for autonomous driving. We follow Eigen *et al.* [10] regarding splitting the training and test sets, where 23488 images containing 32 scenes are used for training, while 697 photos of the remaining 29 scenes are used for evaluation. We evaluate and compare the network based on its performance on top of this training set and test set. Our previous work set the prediction range to [0-80m] and put all the pixel points with depth values greater than 80m to 80m.

D. NYU DEPTH V2 DATASETS

The NYU Depth V2 [18] dataset was created by New York University. The creators used Microsoft's Kinect depth camera to record various indoor scenes. The NYU Depth V2 dataset contains about 120,000 RGB images and their corresponding depth maps. We trained 24,231 image pairs from 249 of these scenes based on previous work and selected an additional 654 images from 215 scenes for testing. We use the aligned depth image pairs provided by the dataset for training and prediction and set the maximum distance to 10m.

E. EVALUATION RESULT

The results of the evaluation on the public dataset KITTI are shown in the table below. Combining the information in the table shows that our model is superior to the methods listed. In the depth range of 0-80m, as shown in Table 1 below, we improved the accuracy by 0.1% in $\delta < 1.25^3$ and made the Sq Rel, RMSE decrease by 0.01. In the depth range of 0-50m, our network improves the accuracy by 0.1% for δ <1.25, δ <1.25^2 and decreases in four metrics Abs Rel, Sq Rel, RMSE, and RMSE log. The results of our method are equal to or better than those of the enumerated method in all metrics. The measurement results on the public NYU Depth V2 dataset are shown in Table 3 below, where we have a 0.1% improvement in $\delta < 1.25^3$ and a 0.01 decrease in the RMSE log metric. In summary, our proposed network can perform a more accurate pixel-level depth estimation. Due to some limitations on the KITTI website for submissions, this algorithm cannot be evaluated and assessed against all current methods at this time. Moreover, the official website of NYU Depth V2 does not have a ranking of all existing methods. Therefore, in this paper, only some representative algorithms with good results are selected for comparison.

TABLE 2. Performance comparison on KITTI Eigen split with a set distance of 0-50m.

Method	cap	higher is better			lower is better			
		δ<1.25	δ<1.25^2	δ<1.25^3	Abs Rel	Sq Rel	RMSE	RMSE *log*
Garg et al. [33]	0-50m	0.740	0.904	0.962	0.169	1.080	5.104	0.273
Godard et al. [34]	0-50m	0.873	0.954	0.979	0.108	0.657	3.729	0.194
Kuznietsov et al. [15]	0-50m	0.875	0.964	0.988	0.108	0.595	3.518	0.179
Gan et al. [58]	0-50m	0.898	0.967	0.986	0.094	0.552	3.133	0.165
DORN [1]	0-50m	0.936	0.985	0.995	0.071	0.268	2.271	0.116
Bts-Resnet101 [4]	0-50m	0.962	0.994	**0.999**	0.058	0.183	1.995	0.090
Bts-Densenet161 [4]	0-50m	0.964	0.995	**0.999**	0.057	0.175	1.949	0.088
MAPNet-Resnet101	0-50m	0.964	**0.996**	**0.999**	0.057	**0.165**	**1.910**	**0.085**
MAPUnet-Densenet161	0-50m	**0.965**	0.995	**0.999**	**0.056**	0.170	1.923	0.086

TABLE 3. Performance comparison on top of the NYU Depth V2 dataset, with a set distance of 0-10m.

Method	cap	higher is better			lower is better		
		δ<1.25	δ<1.25^2	δ<1.25^3	Abs Rel	RMSE	RMSE *log*
Make3d [6]	0-10m	0.447	0.745	0.897	0.349	1.214	-
Liu et al. [12]	0-10m	0.650	0.906	0.976	0.213	0.759	0.087
Eigen et al. [10]	0-10m	0.769	0.950	0.988	0.158	0.641	-
Chakrabarti et al. [59]	0-10m	0.806	0.958	0.987	0.149	0.620	-
Qi et al. [60]	0-10m	0.834	0.960	0.990	0.128	0.569	0.057
Yin et al. [56]	0-10m	0.875	0.976	0.994	0.108	0.416	0.048
Bts [4]	0-10m	0.885	0.978	0.994	0.110	0.392	0.047
DenseDepth [28]	0-10m	**0.895**	**0.980**	0.996	**0.103**	**0.390**	0.043
MAPUNet-Resnet101	0-10m	0.874	0.976	0.996	0.122	0.405	0.050
MAPUnet-Densenet161	0-10m	0.888	0.979	**0.997**	0.109	0.393	**0.040**

TABLE 4. Study of ablation experiments using the KITTI dataset (the depth range: [0-80]m). Baseline: represents the network with ResNet-101 as the backbone network only, without adding the dense cavity pyramid module and without densely connected decoders. A: represents the network with the dense atrous pyramid module added. D: represents the network with densely connected decoders. All variants set the loss function such that $\lambda = 0.9$ in Equation 4.

Variant	Params (M)	higher is better			lower is better				
		δ<1.25	δ<1.25^2	δ<1.25^3	SiLog	Abs Rel	Sq Rel	RMSE	RMSE *log*
Baseline	59.86	0.941	0.991	0.998	9.6859	0.0705	0.310	3.258	0.109
Baseline +A	56.17	0.952	0.992	0.998	9.2432	0.0620	0.256	2.828	0.100
Baseline +A+D	60.89	**0.955**	**0.993**	0.998	**8.9984**	0.0620	0.250	**2.708**	0.097
MAPUnet-Densenet	40.37	**0.955**	0.992	**0.999**	9.0550	**0.0610**	**0.242**	2.741	**0.096**

F. ABLATION STUDY

In this section, we evaluate different network variants and analyze the reasons that affect network performance. We use the ResNet-101 network as the backbone network. Based on a Unet++-like network structure, we densely connect the decoders of each stage and add a dense atrous pyramid

TABLE 5. The experimental results of the dense atrous pyramid with different combinations of radii under the KITTI dataset are as follows (the depth range: [0-80]m). The corresponding node positions are the same as those in 3.2 above, i.e. $\left(X^{2,1}, X^{1,1}, X^{0,1}, X^{1,2}, X^{0,2}, X^{0,3}\right)$. We also use the ResNet-101 network as the backbone network.

Variant	Params (M)	higher is better			lower is better				
		δ<1.25	δ<1.25^2	δ<1.25^3	SiLog	Abs Rel	Sq Rel	RMSE	RMSE *log*
DASPP(3,3,6,3,12,18)	60.89	0.931	0.990	0.998	10.1057	0.0730	0.330	3.258	0.109
DASPP(3,3,12,6,18,24)	60.89	0.937	0.991	0.999	9.9819	0.0739	0.309	3.219	0.112
DASPP(3,6,12,6,18,24)	60.89	0.942	0.989	0.998	9.6704	0.0666	0.278	2.933	0.106
DASPP(3,6,3,12,18,24)	60.89	0.950	0.992	0.998	9.3378	0.0633	0.262	2.850	0.101
DASPP(3,6,12,18,24,30)	60.89	0.945	0.992	0.998	9.6607	0.0670	0.270	2.861	0.104
DASPP(3,6,6,12,18,24)	60.89	**0.955**	**0.993**	**0.998**	**8.9984**	**0.0620**	**0.250**	**2.708**	**0.097**

TABLE 6. Experimental results with different U-shaped network structures under the KITTI dataset (the depth range: [0-80]m). We use the ResNet-101 network as the backbone network.

Variant	Params (M)	higher is better			lower is better				
		δ<1.25	δ<1.25^2	δ<1.25^3	SiLog	Abs Rel	Sq Rel	RMSE	RMSE *log*
Unet	54.88	0.946	0.992	**0.998**	9.7362	0.0683	0.283	3.074	0.106
Unet++	56.17	0.952	0.992	**0.998**	9.2432	0.0620	0.256	2.828	0.100
Unet3+	56.60	0.950	0.991	**0.998**	9.2607	0.0670	0.301	2.917	0.106
MAPUnet	60.89	**0.955**	**0.993**	**0.998**	**8.9984**	**0.0620**	**0.250**	**2.708**	**0.097**

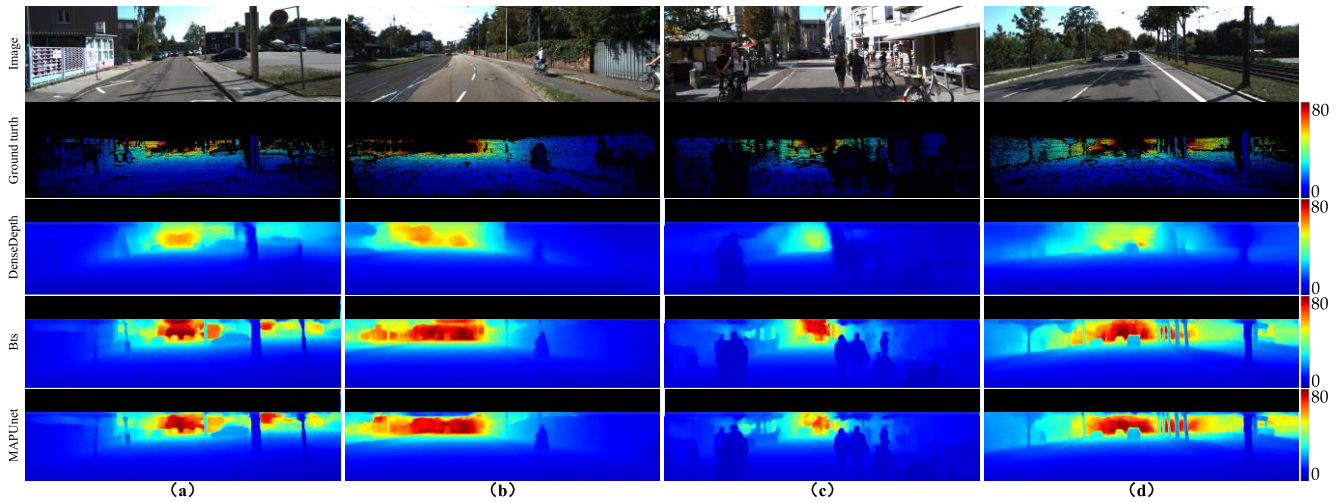

FIGURE 6. Qualitative results of KITTI Eigen test splitting, where the units of the numbers in the pseudo-color scale bar are meters (m). The proposed method yields more precise boundaries from vehicles, traffic signs, and pedestrians and estimates more accurate depth information.

module to explore the impact of the added factors on the network performance. Combining the following Table. 4 shows that the network's performance steadily improves as the core factors increase and that the operation of the dense atrous pyramid module and the dense connection decoder have comparable effects on the network. Although the network increases the training parameters by 4.7M, it achieves better depth estimation and reaches a more advanced level.

G. EXPERIMENTS WITH COMBINATION OF DIFFERENT ATROUS CONVOLUTIONS

Dense atrous pyramids using different combinations of sizes enhance the performance of the network structure proposed in this paper, so we would like to gain insight into the performance impact of other radius combinations of atrous pyramid layers on MAPUnet networks. We select r = 3, r = 6, r = 12, r = 18, r = 24, r = 30, these six values as the

TABLE 7. Network parameters and running time of the proposed network and the compared networks.

Network Name	Params(M)	KITTI Time(s)	NYU Time(s)
DenseDepth	44.32	1.789	1.698
Bts	68.5	0.063	0.065
MAPUnet-resnet101	60.89	0.085	0.057
MAPUnet-densenet161	40.37	0.086	0.059

chosen values of the atrous radius, and test them on the KITTI dataset. When choosing the radius of the atrous convolution, we follow the principle that the equivalent atrous radius K value at each level should be less than the shortest edge length of the feature map. Otherwise a lot of useless information will be introduced into the network. Since the size of the feature map becomes more significant as the network level becomes shallower, the selection of the atrous radius also follows the principle of varying from small to large in the transducer from deep to shallow levels. According to the experimental data in Table. 5, there is some improvement in the effectiveness of the network as the size of the atrous radius used becomes larger. Still, the effect regresses when either the chosen atrous radius becomes too large or too small. Also, it can be seen that the effect of choosing a smaller radius of the atrous convolution layer is better than that of choosing a larger radius of the atrous convolution layer in the first node of the transducer-connected encoder. In the same network framework, choosing the right combination of radii for the atrous convolution pyramid can improve the accuracy by more than 2% (in terms of $\delta < 1.25$).

H. EXPERIMENTS WITH DIFFERENT TRANSDUCER-DECODERS

This paper also conducts experiments for networks with different decoders. While keeping the encoder as ResNet-101, we conducted experiments on the KITTI dataset using a network structure similar to Unet, Unet++, and Unet3+, respectively. Although Unet3+ achieves the best segmentation results in the field of medical segmentation, the Unet3+ structure does not perform as well as the Unet++ structure in terms of depth estimation, as shown in Table. 6 below. The reason is that although both depth estimation and image segmentation are pixel-level computer vision tasks, the depth values change incrementally throughout the image. Therefore the brute force merging of all scales together and then decoding step by step from deep to shallow networks is not conducive to the network forming a robust function for solving 3D depth information from two dimensions. Therefore, we adopt the structure of a densely connected decoder based on Unet++, and the experimental results show a steady improvement in most of the metrics.

I. QUALITATIVE RESULT

Finally, we qualitatively discuss the results of our work and that of the competition. As seen in Fig. 6 (a), (b), (c), and (d), our work shows more accurate object boundaries, and the outlines of signs, people, and vehicles on some poles at a long distance can be estimated more accurately. However, in the test results output by the network on the KITTI dataset,

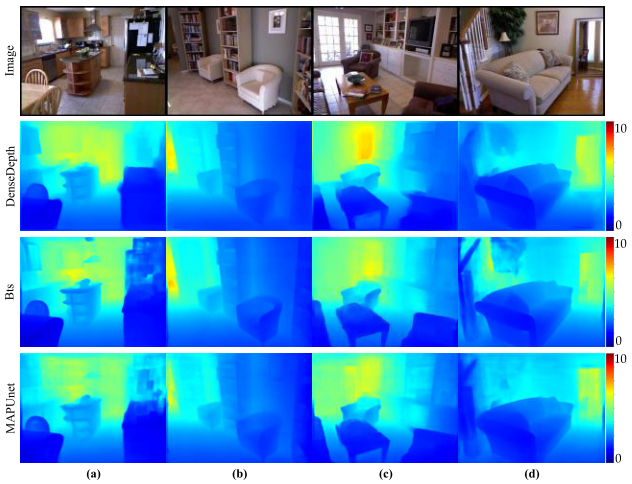

FIGURE 7. Qualitative results in the NYU Depth V2 dataset, where the units of the numbers in the pseudo-color scale bar are meters (m). The proposed method more accurately recovers the depth information of the image, such as the distant bookshelf outline in Figure (b).

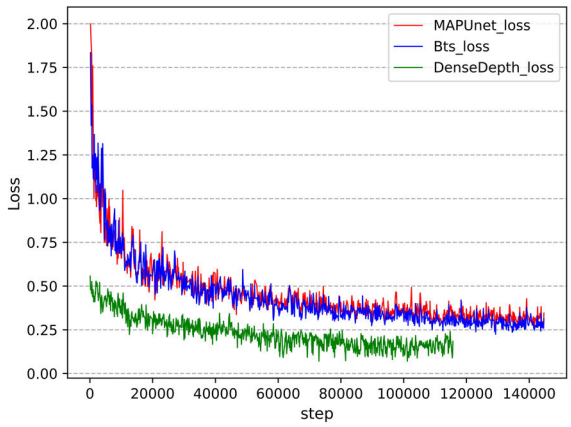

FIGURE 8. The loss convergence curves of the compared algorithms are plotted. With the same KITTI dataset, the Bts and MAPUnet networks were trained for 25 epochs with a batch of 2. The DenseDepth network was trained for 20 epochs with a batch of 2.

the images have some artifacts in either the sky or the upper regions of the scene. We believe this is caused by the very sparse ground truth depth data. Because some image regions lack valid depth values in the entire dataset, it is impossible to train the network appropriately for these regions. The test results on the NYU Depth V2 dataset are shown in Fig. 7 below. The MAPUnet network can predict the contours of objects in more distant regions, such as bookshelves, glass wooden doors, etc.

Also, in Table. 7, the number of model parameters and the running time of the MAPUnet network proposed in this paper, and the compared algorithms are listed. Among

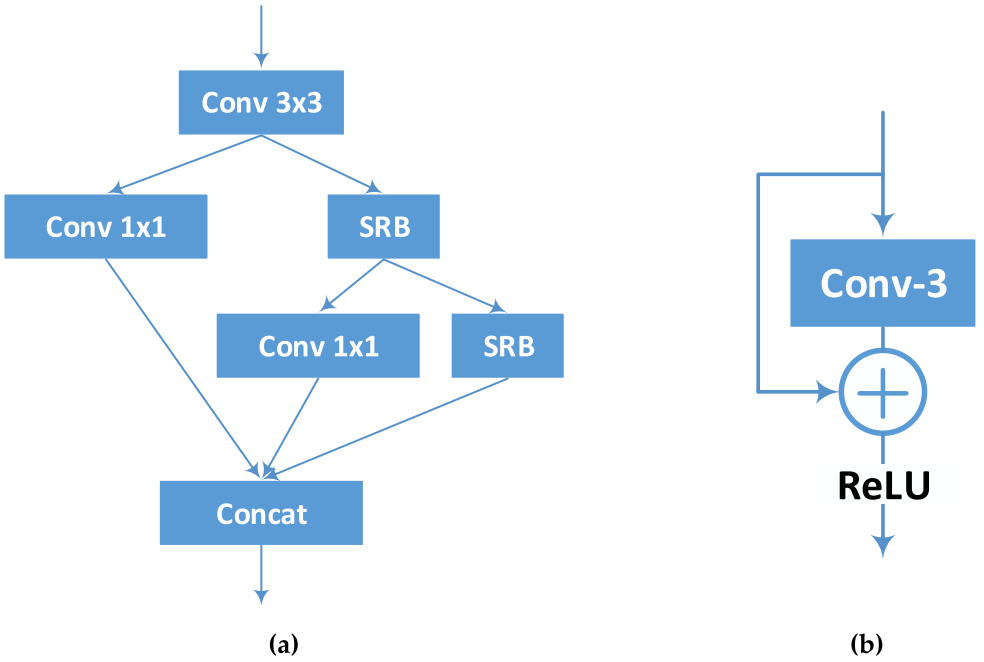

FIGURE 9. Figure (a) shows the module used in [61] instead of the convolutional layer, using deep residual connections to improve the feature extraction capability of the network. Figure (b) shows the SRB module in Figure (a). The segmentation effect is enhanced by more than three percentage points in [61].

them, MAPUnet-resnet101 has FLOPs of 1.59×10^{11}, and MAPUnet-densenet161 has FLOPs of 1.52×10^{11}. We also present the convergence curves about losses, as shown in Figure 8. Our network and the Bts network are trained for 25 epochs, while DenseDepth is trained for 20 epochs as described in its paper. As shown in Figure 8, all three loss curves show a decreasing trend, indicating that the network is converging. In contrast, the DenseDepth network's loss curve values have been smaller because DenseDepth uses a loss function that is different from the MAPUnet network and Bts network.

V. DISCUSSION

We also find some problems during the research process, which are briefly stated here as follows.

- We add the FDB module [61] to the decoder layer of the network, i.e., the convolutional layer is replaced with the FDB module in the decoder stage, and the FDB module is shown in Fig. 9(a) below. The test results show that it does not improve the network's performance but degrades the network's performance. The reason is that the modified network structure is too deep in layers, so it is difficult to propagate the gradient to the deepest layer of the network.
- Although the window part is transparent in color, the estimated depth of the window part should be approximate to the depth of the window frame when the window is not lowered normally. Still, the test result is the opposite, as shown in Fig. 10 below. Observing the ground truth plot, we can see that the radar signal received at the window position is weak, so the network cannot be trained well. This part of the depth estimation task can be improved by combining semantic segmentation information.

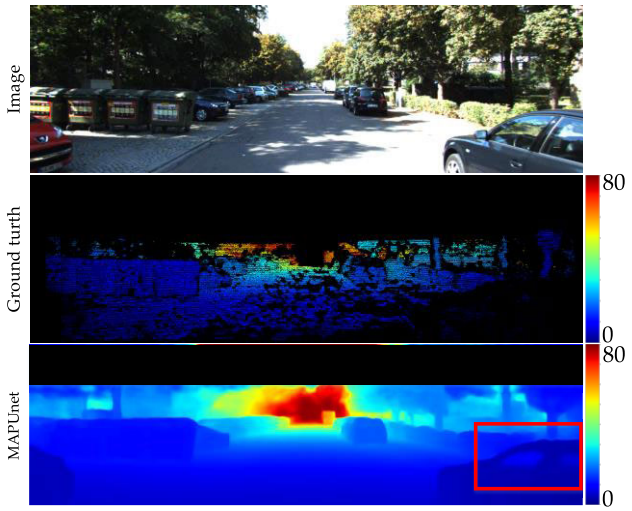

FIGURE 10. The outline of the car frame is visible in the lower right corner of the depth image estimated in this paper, while the windows of the car are not rolled down according to the input image. The units of the numbers in the pseudo-color scale bar are meters (m).

VI. CONCLUSION

This paper proposes a monocular depth estimation network MAPUnet with supervised mode, which achieves better test results. Using some achievements in other fields of deep learning, we design an encoder-transducer-decoder-based neural network architecture, based on the Unet++ framework, densely connected decoders. We introduce a dense atrous pyramid structure in the transducer phase to enable the network to progressively deeper network information layer by layer. Testing and validation on the public KITTI dataset and the NYU Depth V2 dataset outperform the other methods

compared to root mean square error and mean absolute log error, which also side-by-side illustrates the higher accuracy of the depth values output by the network. Although we achieve good depth estimation results, there are still some problems, such as the existence of artifacts in humans and vehicles, and the inability to estimate the depth of the windows, and the high sparsity of the data is the main reason for the problem. Therefore, future work is planned to apply the network structure in unsupervised depth estimation for more reasonable utilization of sparse ground truth data values, combined with the information of semantic segmentation or using rank loss (rank loss) to improve the network performance further.

REFERENCES

[1] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2002–2011.
[2] *T. Autopilot*. Accessed: Dec. 19, 2019. [Online]. Available: https://www.tesla.com/autopilot
[3] *U. ATG*. Accessed: Dec. 19, 2019. [Online]. Available: https:www.uber.com/us/en/atg
[4] J. H. Lee, M.-K. Han, D. W. Ko, and I. H. Suh, "From big to small: Multi-scale local planar guidance for monocular depth estimation," 2019, *arXiv:1907.10326*. [Online]. Available: https://arxiv.org/abs/1907.10326
[5] I. P. Howard, *Perceiving in Depth Volume 1 Basic Mechanisms* (Psychophysics and Analysis). New York, NY, USA: Oxford Univ. Press, 2012.
[6] A. Saxena, M. Sun, and A. Y. Ng, "Make3D: Learning 3D scene structure from a single still image," in *Proc. IEEE 11th Int. Conf.*, May 2008, pp. 824–840.
[7] A. Saxena, S. H. Chung, and A. Y. Ng, "Learning depth from single monocular images," in *Proc. NIPS*, 2005, pp. 1–8.
[8] K. Karsch, C. Liu, and S. B. Kang, "Depth transfer: Depth extraction from video using non-parametric sampling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 11, pp. 2144–2158, Nov. 2014.
[9] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 42, pp. 7–42, Apr. 2002.
[10] 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: https://arxiv.org/abs/1406.2283
[11] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," in *Proc. Int. Conf. Comput. Vis.* Santiago, Chile: CentroParque Convention Center, 2015.
[12] 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.
[13] 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.
[14] 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. 2018.
[15] Y. Kuznietsov, J. Stuckler, and B. Leibe, "Semi-supervised deep learning for monocular depth map prediction," in *Proc. IEEE Conf. Comput. Pattern Recognit.*, Jul. 2017, pp. 6647–6655.
[16] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3684–3692.
[17] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *Proc. Int. Conf. Learn. Represent.* San Juan, Puerto Rico: Caribe Hilton, 2016.
[18] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. 12th Eur. Conf. Comput. Vis.* Berlin, Germany: Springer, 2012, pp. 746–760.
[19] M. H. Baig and L. Torresani, "Coupled depth learning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2016, pp. 1–10.
[20] R. Ranftl, V. Vineet, Q. Chen, and V. Koltun, "Dense monocular depth estimation in complex dynamic scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4058–4066.
[21] R. Furukawa, R. Sagawa, and H. Kawasaki, "Depth estimation using structured light flow–analysis of projected pattern flow on an object's surface," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4640–4648.
[22] W. Zhuo, M. Salzmann, X. He, and M. Liu, "Indoor scene structure analysis for single image depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 614–622.
[23] F. Zhang, V. Prisacariu, R. Yang, and P. H. S. Torr, "GA-Net: Guided aggregation net for end-to-end stereo matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 185–194.
[24] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1119–1127.
[25] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3917–3925.
[26] 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.*, Jul. 2017, pp. 5354–5362.
[27] I. Alhashim and P. Wonka, "High quality monocular depth estimation via transfer learning," 2018, *arXiv:1812.11941*. [Online]. Available: https://arxiv.org/abs/1812.11941
[28] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
[29] S. Aich, J. M. U. Vianney, A. Islam, M. Kaur, and B. Liu, "Bidirectional attention network for monocular depth estimation," 2020, *arXiv:2009.00743*. [Online]. Available: https://arxiv.org/abs/2009.00743
[30] Z. Xia, P. Sullivan, and A. Chakrabarti, "Generating and exploiting probabilistic monocular depth estimates," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 65–74.
[31] K. Xian, J. Zhang, O. Wang, L. Mai, Z. Lin, and Z. Cao, "Structure-guided ranking loss for single image depth prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 611–620.
[32] 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.*, 2016, pp. 740–756.
[33] C. Godard, O. M. Aodha, and G. J. Brost, "Unsupervised monocular depth estimation with left-right consistency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 270–279.
[34] 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.
[35] F. Aleotti, F. Tosi, M. Poggi, and S. Mattoccia, "Generative adversarial networks for unsupervised monocular depth prediction," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Munich, Germany, 2019, pp. 337–354.
[36] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3D packing for self-supervised monocular depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2485–2494.
[37] H. Zhan, R. Garg, C. S. Weerasekera, K. Li, H. Agarwal, and I. Reid, "Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 340–349.
[38] V. Casser, S. Pirk, R. Mahjourian, and A. Angelova, "Unsupervised monocular depth and ego-motion learning with structure and semantics," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2019, pp. 381–388.
[39] 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.
[40] A. Eldesokey, M. Felsberg, K. Holmquist, and M. Persson, "Uncertainty-aware CNNs for depth completion: Uncertainty from beginning to end," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 12014–12023.
[41] M. Vankadari, S. Garg, A. Majumder, S. Kumar, and A. Behera, "Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 443–459.
[42] R. Cheng, R. Razani, Y. Ren, and L. Bingbing, "S3Net: 3D LiDAR sparse semantic segmentation network," 2021, *arXiv:2103.08745*. [Online]. Available: https://arxiv.org/abs/2103.08745

[43] M. Klingner, J. A. Termöhlen, J. Mikolajczyk, and T. Fingscheidt, "Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 582–600.
[44] 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.
[45] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[47] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: Redesigning skip connections to exploit multiscale features in image segmentation," *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Jun. 2020.
[48] H. Huang, L. Lin, R. Tong, H. Hu, Q. Zhang, Y. Iwamoto, X. Han, Y.-W. Chen, and J. Wu, "UNet 3+: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1055–1059.
[49] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *Int. J. Robot. Res.*, vol. 32, no. 11, pp. 1231–1237, Sep. 2013.
[50] M. Holschneider, R. Kronland-Martinet, J. Morlet, and P. Tchamitchian, "A real-time algorithm for signal analysis with the help of the wavelet transform," in *Wavelets*. Berlin, Germany: Springer-Verlag, 1989, pp. 286–297.
[51] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.
[52] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, and T. Killeen, "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8026–8037.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[54] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 764–773.
[55] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[56] W. Yin, Y. Liu, C. Shen, and Y. Yan, "Enforcing geometric constraints of virtual normal for depth prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5684–5693.
[57] Z. Zhang, S. Lathuiliere, E. Ricci, N. Sebe, Y. Yan, and J. Yang, "Online depth learning against forgetting in monocular videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4494–4503.
[58] Y. Gan, X. Xu, W. Sun, and L. Lin, "Monocular depth estimation with affinity, vertical pooling, and label enhancement," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 224–239.
[59] A. Chakrabarti, J. Shao, and G. Shakhnarovich, "Depth from a single image by harmonizing overcomplete local network predictions," 2016, *arXiv:1605.07081*. [Online]. Available: http://arxiv.org/abs/1605.07081
[60] 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.
[61] R. Bai, S. Jiang, H. Sun, Y. Yang, and G. Li, "Deep neural network-based semantic segmentation of microvascular decompression images," *Sensors*, vol. 21, no. 4, p. 1167, Feb. 2021.

YIFAN YANG received the B.S. degree in automation engineering from Harbin Engineering University (HEU), Harbin, China, in 2017. She is currently pursuing the Ph.D. degree in mechanical and electronic engineering with Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun, China. Her research interests include depth prediction, depth completion, and simultaneous localization and mapping.

YUQING WANG received the B.S. and M.S. degrees in electrical engineering from Jilin University, Changchun, China, in 2002 and 2005, respectively, and the Ph.D. degree in optical engineering from Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun, in 2008. From August 2008 to August 2010, he was a Postdoctoral Fellow in mechanical and electronic engineering with CIOMP, where he is currently an Associate Professor. His research interests include image fusion, image quality assessment, small target tracking, and image enhancing.

CHENHAO ZHU received the B.S. degree in mechanical engineering from Beijing Institute of Technology (BIT), Beijing, China, in 2017. He is currently pursuing the Ph.D. degree in mechanical and electronic engineering from Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun, China. His research interests include spacecraft navigation, guidance, and control systems.

MING ZHU is currently a Research Fellow and the Supervisor of Ph.D. Candidates of Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His research interests include digital image processing, television tracking, and automatic target recognition technology.

HAIJIANG SUN was born in Huinan, China, in 1981. He received the Ph.D. degree in electronic engineering from Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, in 2010. He is currently a Professor with the Perception and Display Laboratory, CIOMP. His research interests include target recognition and tracking technology and high definition video image enhancement display. He is an Associate Editor of the *Liquid Crystals*, *Virtual Reality*, and *Optics and Precision Engineering*.

TIANZE YAN received the B.S. degree in measurement and control technology and instrument major from Changchun University of Science and Technology (CUST), Changchun, China, in 2019. He is currently pursuing the Ph.D. degree in mechanical and electronic engineering with Changchun Institute of Optics, Fine Mechanics and Physics (CIOMP), Chinese Academy of Sciences, Changchun. His research interests include three-dimensional reconstruction, depth prediction, and depth completion.