Information Fusion 58 (2020) 1–12

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full Length Article

IVFuseNet: Fusion of infrared and visible light images for depth prediction

Yuqi Li, Haitao Zhao*, Zhengwei Hu, Qianqian Wang, Yuru Chen

East China University of Science and Technology, 130 Meilong Road, Xuhui District, Shanghai 200237, PR China

ARTICLE INFO

Keywords:
Depth prediction
Partially coupled filter
Adaptive weighted fusion
Visible image
Infrared image

ABSTRACT

Depth prediction is an essential component in the research of unmanned driving. Most existing research works predict depth only based on visible light images or infrared images. However, both visible light images and infrared images have their own advantages and disadvantages, and these two kinds of images contain complementary information when the images are filmed from the same scence. In order to fuse the complementary information and predict depth under various conditions, this paper proposes a convolutional-neural-network-based architecture, called infrared and visible light images fusion network (IVFuseNet), for depth prediction. Specifically, we construct common-feature-fusion subnetwork, full-feature-fusion subnetwork, and high-resolution reconstruction subnetwork, aiming to leverage the complementarity of these two kinds of images. The common-feature-fusion subnetwork adopts a two-stream multilayer convolutional structure whose filters for each layer are partially coupled to fuse the common features extracted from infrared images and visible light images respectively. The full-feature-fusion subnetwork fuses the two-stream features generated from the common-feature-fusion subnetwork by adaptive fusion weights instead of prefixed fusion weights. Additional, residual dense convolution that can accurately map the fused low-resolution features to the corresponding high-resolution features is adopted in the high-resolution reconstruction subnetwork to enhance the reconstruction of the details for depth prediction. All three subnetworks collaborate together to conduct the depth prediction task. Our *NUST-SR* dataset is composed of the actual road scenes captured while unmanned vehicle driving. The proposed IVFuseNet obtains the best performances on this dataset. IVFuseNet decreases the root mean squared error to 3.4513 and the mean relative error to 0.1651 respectively and outperforms other methods. The model and dataset are available at https://github.com/liyuqi1234/IVFN.

1. Introduction

The vision-based unmanned driving has attracted a lot of research in both academia and industry [1–5]. The depth prediction is one of the main contents in this researching field, studying the methods to restore the depth information of scenery from two-dimensional images. The depth information can help understand the geometric relationship in the scene and judge the distance between the current vehicle and other objects, and further enhance the safety of unmanned driving.

Most research works predict depth only based on visible light images. Traditional methods for depth prediction are usually required to construct hand-crafted features, such as scale-invariant feature transform (SIFT) and the histogram of oriented gradient (HOG) [6–8]. Recently, methods based on convolutional neural network (CNN) can automatically extract more effective features to predict the depth and achieve substantial gains [9–13]. The multi-scale deep network [9] is the first CNN-based method to predict depth with visible light images. Other CNN-based methods combined with condition random field (CRF) to obtain depth maps [14,15]. For example, Li et al. [42] used a deep convolutional neural network (DCNN) to get depth maps of visible light images and then used CRF to refine the depth maps.

However, visible light images are restricted by the application scenarios in depth prediction. For example, there are no reliable features that can be extracted in very low light level conditions such as overcast night or in degraded visibility conditions if only visible light images are available. In contrast, thermal infrared images detect the thermal radiation of objects and convert it into images, this make infrared images have more advantages. First, infrared images are not affected by brightness changing, and can extract stable features under the conditions of intense change of light and weak light at night. Second, infrared images have high penetration and have the ability to penetrate rain, fog and snow, therefore, they solve the problem caused by occlusion. Third, the detection distance of infrared images is much further than that of visible light images. Very distant objects in the infrared image are also clear.

Due to the above reasons, thermal-infrared-image-based depth prediction methods have been proposed [16–19]. Xi et al. [16] extracted the features based on multi-scale and spatial context information to make up for the insufficient local information of infrared images and selected

* Corresponding author.
E-mail addresses: Y30170645@mail.ecust.edu.cn (Y. Li), haitaozhao@ecust.edu.cn (H. Zhao), Y30170632@mail.ecust.edu.cn (Y. Chen).

https://doi.org/10.1016/j.inffus.2019.12.014
Received 3 February 2019; Received in revised form 9 October 2019; Accepted 25 December 2019
Available online 26 December 2019

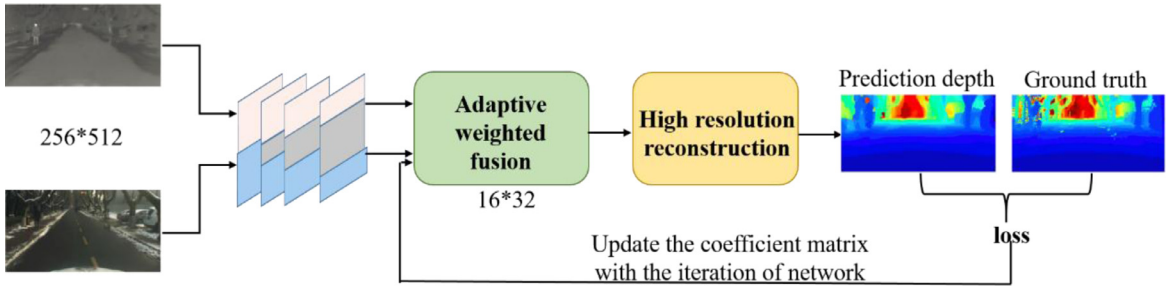

Fig. 1. Diagram of the proposed IVFuseNet. The pink blocks represent the stream of infrared images and the blue blocks represent the stream of the visible images. The gray blocks are designed to extract fused features from both infrared and visible light images.

the features suitable for depth prediction of infrared images by independent components analysis (ICA), then predicted the depth through nonlinear support vector regression (SVR). Sun et al. [18] proposed a method based on kernel principal component analysis (KPCA) and a fully connected neural network. In their work, KPCA was used to extract nonlinear features suitable for infrared images, then the features and the corresponding ground-truth depths were used to train a BP neural network.

Although the infrared images make up the shortcomings of visible light images in some respects, the performance of depth prediction based on infrared images alone is not accurate enough for small objects and the edge of objects since infrared images lack the texture information and have low contrast. Visible light images have rich texture information and high contrast which make up the shortcomings of infrared images. Through the above analysis, the depth prediction based on the fusion of infrared and visible light images may achieve better performances than the depth prediction only based on visible light images or infrared images.

In this paper, we leverage the fusion information of infrared and visible light images to predict depth. Please note that registration of two kinds of images is a critical issue in image fusion tasks [41,42]. Due to the infrared and visible light images in the dataset we used are already registered, we do not consider the image registration. Many successful methods for image fusion have been proposed [20–27]. Laplacian pyramid [20] and wavelet transform [21] were widely adopted for image fusion. In recent works [22–24,27], CNN-based methods have been proposed to fuse different kinds of images. For example, Prabhakar et al. [24] proposed exposure fusion with extreme exposure image pairs, CNN was used to fuse the luminance channel of the exposure image pairs and the weighted fusion strategy was used to fuse the chrominance channels of the input images.

Infrared images and visible light images convey different information as to the depth prediction. Predicting the depth of the low light scenes, infrared images contribute more than visible light images. When we predict the depth of objects which is full of texture and under adequate illumination, the result is reversed. The above-mentioned fusion methods generally fuse infrared images and visible light images with an equal-weight or prefixed hand-crafted weights, neglecting the varying contributions of the two kinds of images on predicting the depth of different objects in different scenes.

In this paper, we propose a CNN-based architecture called IVFuseNet (Fig. 1), which adaptively fuses the complementary information of infrared and visible light images, for depth prediction, endowed with three subnetworks: the common-feature-fusion subnetwork, the full-feature-fusion subnetwork, and the high-resolution reconstruction subnetwork.

First, we considered that infrared and visible light images make different contributions for depth prediction in different scenes. We proposed the adaptive weighted fusion strategy, in which the coefficient matrix represents the different contributions of two kinds of images. In order to keep significant features we proposed partial coupled filters in the common-feature-fusion subnetwork to further enhance the features before adaptive weighted fusion. The coupled filters help discover more discriminative features which are prone to be overlooked from each other images.Furthermore, resolution of the fused features is decreased during the convolutions and smaller than the resolution of the ground truth depth images. We need to recover the resolution of the fused feature before the depth prediction. Inspired by [28,29], the residual dense convolution (RDC) has good performance in super-resolution reconstruction. So in the high-resolution reconstruction subnetwork, we adopt RDC in the high-resolution reconstruction subnetwork, which can also enhance the fused features. Our contributions are as follows:

- The partially coupled filters are designed in common-feature-fusion subnetwork to extract features from infrared and visible light images and learn the transferable features between infrared images and visible light images. That is equivalent to retain the individual features of two types of input images while fusing common features in the process of subnetwork learning.
- The adaptive weighted fusion method considers the different contributions of infrared and visible light images in different scenes and the adaptive fusion coefficient matrix is trained by the full-feature-fusion subnetwork without the prefixed hand-crafted weights.
- The residual dense convolution blocks which help recover finer details and improve the effect of feature fusion are applied in the high-resolution subnetwork for the task of depth prediction. The high-resolution subnetwork maps the low-resolution features to the corresponding high resolution features accurately for supervised learning.

2. Related work

Directly related to our work are some methods that predict the depth and fuse infrared and visible light images. CNN-based methods have already become the most important methods for depth prediction. Eigen et al. [10] proposed a three-scale convolutional network to predict depth. The output prediction of each scale was passed to the next scale in their work. Three scales were used to refine the predicted depth progressively and extract more image details. Gu et al. [19] designed a 2D residual neural network and a 3D CNN combined network, which considered information between two consecutive frames of images by combining optical flow information. Instead of using fully-connected layers at last, Laina et al. [30] predicted depth with a fully convolutional residual network and presented a novel up-sampling block to improve the output resolution. These methods hardly make full use of all hierarchical information of the original low-resolution features. Inspired by [28,29], we adopt residual dense convolution blocks in our network. The residual dense convolution blocks allow connections from the output of every preceding layer to the current layer directly to leverage all hierarchical features.

When the unmanned vehicle drives under different light conditions, we might not be able to obtain sufficient information if we only use images acquired from a single RGB camera. Our method predicts depth by

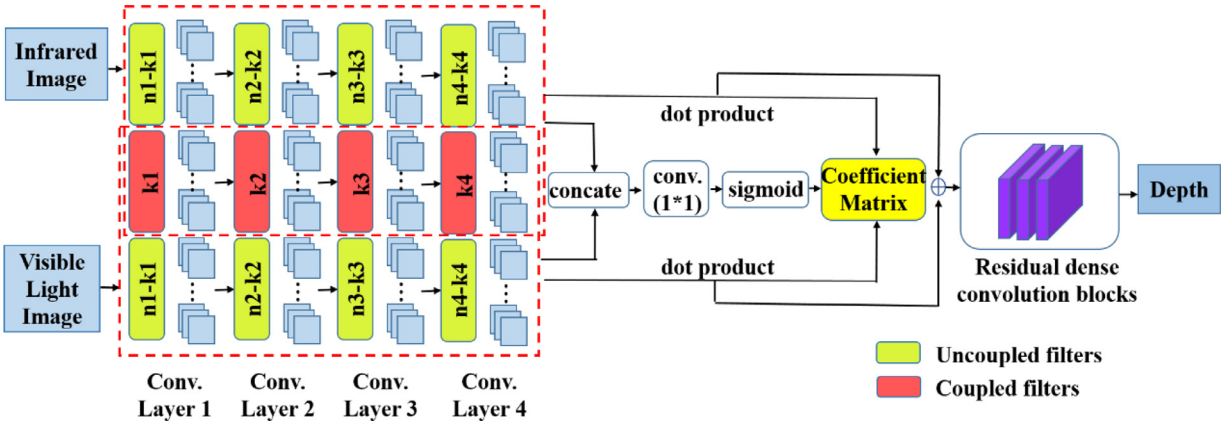

Fig. 2. Diagram of our IVFuseNet. The top red dotted frame presents the stream of infrared images, the bottom red dotted frame presents the stream of visible light images and the overlap is the coupled filters of two kinds of images. The features of infrared and visible light images are then fused by the method of adaptive weighted fusion. Finally, residual dense convolution is applied to enhance features and reconstruct high-resolution features.

combining both the information of visible light images and infrared images. Integrating complementary information of visible light images and infrared images is widely studied in the image processing domain [25–27,31]. Li et al. [22] designed a DenseFuse network to get more meaningful features from source images and fused the features with equal-weight. Ma et al. [27] proposed a generative adversarial network called FusionGAN to fuse infrared and visible light images. The generator generated an image, which is fused by infrared information and additional visible information. The discriminator made the fused image have more texture information of visible light images. These methods mainly focused on feature extraction, while the fusion of different features was simply by prefixed hand-crafted weights. However, the weights of the two kinds of images should be variable under different light conditions for depth prediction.

Different from the above fusion methods, first, we design a common-feature-fusion subnetwork with partially coupled filters to fuse common features. Second, we use the strategy of adaptive weighted fusion to fuse the full features from common-feature-fusion subnetwork, this strategy makes infrared images do more contributions in low-light conditions and visible light images do more contributions in sufficient-light conditions for depth prediction.

3. Technical approach

In this section, we describe our infrared and visible light images fusion network (IVFuseNet) for depth prediction. The diagram of our network is shown in Fig. 2. The partial-coupled filters are designed in the common-feature-fusion subnetwork to extract shallow-layer features and fuse the common features of two different kinds of images. In the full-feature-fusion subnetwork, we proposed an adaptive weighted fusion strategy to fuse the full features extracted by the common-feature-fusion subnetwork. Subsequently, we design a high-resolution reconstruction subnetwork to enhance feature details and predict the high-resolution depth map.

3.1. Common-Feature-Fusion subnetwork

As mentioned previously, images acquired from a single RGB camera may contain insufficient information to extract for depth prediction under certain light conditions. In the proposed IVFuseNet, both infrared images and visible light images are used to leverage the complementarity of these two kinds of images. We make infrared images and visible light images as “auxiliary variables” to each other. Because each infrared-visible light image pair represents the same scene. We assume that there exist common features of these two kinds of images. Wang et al. [32] proposed that certain similar features could be transferred from the source domain to the target domain utilized coupled filters. However, not all the features of infrared and visible light images can be represented and transferred to each other even after highly sophisticated operations. In this paper, the features which can be transferred to each other are called “common features” while the features which cannot be transferred to each other are called “individual features”. Based on the above analysis, we design the common-feature-fusion subnetwork to fuse common features firstly.

The detailed mechanism of our common-feature-fusion subnetwork is illustrated by red dotted frames in Fig. 2. We adopt a two-stream structure. For each stream, we use AlexNet [33] as our baseline due to the number of parameters of AlexNet is relatively small. The specific content of common-feature-fusion subnetwork is given in Table 1. Moreover, we can also choose other networks as baseline such as VggNet [34]. This subnetwork takes an infrared image and its corresponding visible light image both with the resolution of 256 × 512 as the input of the subnetwork. Different from traditional two-stream CNN, we design partial coupled filters in each convolution layer to learn the transferable features between infrared and visible light images. The coupled ratios are shown in Table 2.

As shown in Fig. 3, the filters of common-feature-fusion subnetwork can be divided into three classes: filters of the infrared images, filters of visible light images and partially coupled filters of both infrared and visible light images. The partially coupled filters are designed to extract features of both infrared and visible light images. As the “auxiliary variables” of visible light images, infrared images can help discover more discriminative features that are uncaptured from the visible light images under low light conditions. Similarly, as the “auxiliary variables” of infrared images, visible light images can help extract more detail information (texture and edge, etc.) from infrared images by partially coupled filters. The uncoupled filters learn individual features of infrared images and visible light images. The ratio of the number of partially

Tabel 1
The specific content of common-feature-fusion subnetwork.

Layers	Kernel	Strides	Channels	Output size	Coupled ratio
Conv1	11 × 11	4 × 4	96	64 × 128	0
Pool1	2 × 2	2 × 2	/	32 × 64	/
Conv2	5 × 5	1 × 1	256	32 × 64	0.25
Pool2	2 × 2	2 × 2	/	16 × 32	/
Conv3	3 × 3	1 × 1	384	16 × 32	0.5
Conv4	3 × 3	1 × 1	384	16 × 32	0.75

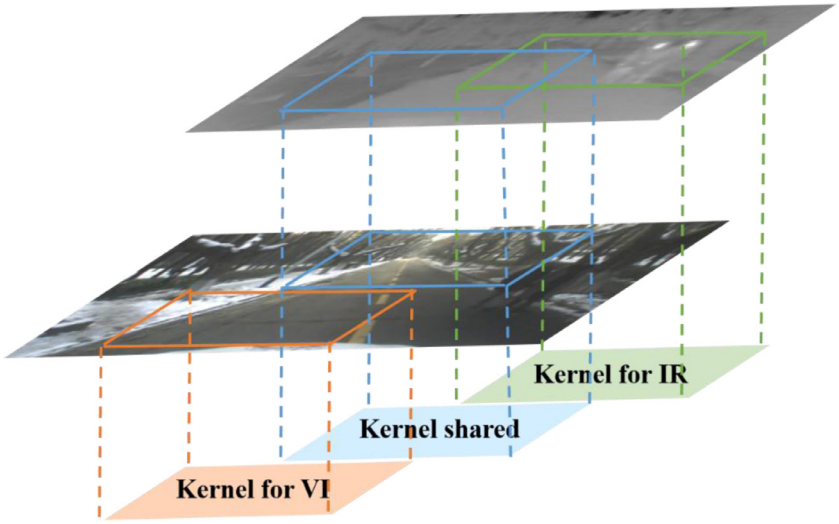

Fig. 3. Kernel for VI and Kernel for IR are uncoupled filters to extract individual features; Kernel coupled are partial coupled filters to extract common features for feature fusion.

Tabel 2
The root mean squared error with different configurations of coupled filters, on the *NUST-SR* dataset.

k1	k2	k3	k4	R1	R2	R3	R4	Rel
0	0	0	0	0	0	0	0	0.171
0	0	0	96	0	0	0	0.25	0.169
0	0	0	192	0	0	0	0.5	0.167
0	0	0	288	0	0	0	0.75	0.167
0	0	0	384	0	0	0	1	0.170
0	0	96	288	0	0	0.25	0.75	0.168
0	0	192	288	0	0	0.5	0.75	0.166
0	0	288	288	0	0	0.75	0.75	0.175
0	64	192	288	0	0.25	0.5	0.75	0.165
0	128	192	288	0	0.5	0.5	0.75	0.177
24	128	192	288	0.25	0.25	0.5	0.75	0.169
96	256	384	384	1	1	1	1	0.171

coupled filters and the number of total filters is called the coupled ratio.

$$R_i = k_i / n_i \ (i = 1, 2, 3, 4) \tag{1}$$

where R_i is the coupled ratio for the i th layer, k_i is the number of partially coupled filters of the i th convolution layer, and n_i is the number of full filters of the i th convolution layer. In this paper, we use the following coupled ratios: 0, 0.25, 0.5, 0.75. We demonstrate the performances of these coupled ratios by a coarse grid search in Section 4.

It can be noticed that the coupled ratios become larger with the increasing of convolutional layers. The analysis shows that shallow convolutional layers extract texture and detail features, which are largely different between infrared images and visible light images. While deeper convolutional layers extract structure and shape features that share common information of both infrared images and visible light images.

In the common-feature-fusion network, we update the filter weights by the backpropagation (BP) method. It can be found that the weights of the coupled filters are updated twice in each training iteration for both the infrared image stream and visible light image stream, and the weights of the uncoupled filters are updated once in each iteration. Therefore, if suppose that we update weights in the infrared stream first and then update weights in the visible light stream, the filter weights are updated as follows:

$$w^n_{infrared} = \begin{cases} w^{n-1}_{ir\_uncoupled} + \mu \frac{\partial \mathrm{L}}{\partial w^{n-1}_{ir_{uncoupled}}} \\ w^{2n-2}_{ir\_coupled} + \mu \frac{\partial \mathrm{L}}{\partial w^{n-1}_{ir\_coupled}} \end{cases} \tag{2}$$

$$w^n_{visible} = \begin{cases} w^{n-1}_{vi\_uncoupled} + \mu \frac{\partial \mathrm{L}}{\partial w^{n-1}_{vi_{uncoupled}}} \\ w^{2n-1}_{vi\_coupled} + \mu \frac{\partial \mathrm{L}}{\partial w^{n-1}_{vi\_coupled}} \end{cases} \tag{3}$$

where n is the iteration numbers, μ is the learning rate, and L is the loss function. The weights of the coupled filters updated in each iteration as follows:

$$\alpha^{2m-1}_{coupled} = \alpha^{2m-2}_{coupled} + \mu \frac{\partial \mathrm{L}}{\partial \alpha^{2m-2}_{coupled}} \tag{4}$$

$$\alpha^{2m-2}_{coupled} = \alpha^{2m-3}_{coupled} + \mu \frac{\partial \mathrm{L}}{\partial \alpha^{2m-3}_{coupled}} \tag{5}$$

In summary, the partially coupled filters are designed in the common-feature-fusion subnetwork to learn the nonlinear transformations of common features of infrared and visible light images, which is equivalent to enhance the features of two kinds of images and fuse the common features.

The features extracted from infrared images and visible light images are divided into common features and individual features. In order to utilize both the common features and individual features for further depth prediction, we propose the full-feature-fusion subnetwork to fuse all features of the two-stream structure. The efficiency of the full-feature-fusion subnetwork will be analyzed in Section 3.2.

3.2. Full-Feature-Fusion subnetwork

After the extraction of the common features by the partially coupled filters, we need to fuse common features and individual features for depth prediction. Due to the different characteristics of the extracted features, it is critical to design an adaptive strategy for feature fusion.

In this paper, we propose an adaptive weighted fusion strategy in our full-feature-fusion subnetwork in Fig. 2. This fusion strategy is composed of three steps. Let $f_{ir} \in R^{b \times w \times h \times c}$ and $f_{vi} \in R^{b \times w \times h \times c}$ denote the features of infrared images and visible light images extracted by the common-feature-fusion subnetwork. Firstly, we concatenate f_{ir} and f_{vi} at the third dimension, this is equivalent to the fused features $f^{fusion} \in R^{b \times w \times h \times 2c}$ of infrared images and visible light images. Secondly, we convolve the fused features with kernel $\mathrm{k} \in R^{2c \times c \times 1 \times 1}$ where 2c is the number of input channels, c is the number of output channels and the kernel size is 1×1. Inspired by [43], due to the fused feature f^{fusion} considers the characteristics of infrared and visible light images at the same time, the output of this convolution operation is related to both two kind of features. Therefore, this procedure learns

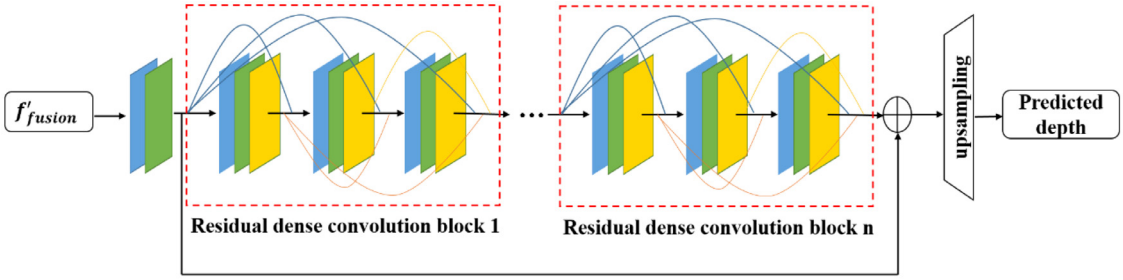

Fig. 4. Each red dotted frame denotes a residual dense convolution block. The blue squares denote convolution, the green squares denote batch normalization and the yellow squares denote activation function.

the correlations of two kinds of features. After getting the initial coefficient matrix M with the same dimension of f_{ir} or f_{vi}, we calculate the dot product of them, which represents their different contributions for depth prediction of different scenes. Thirdly, we design a sigmoid layer to convert each element of M to a form of probability between 0 and 1. After these three steps, we obtain the final coefficient matrix G. The process is as follows:

$$M_{b,i,j,n} = \sum_{n=1}^{2c} f_{b,i,j,n}^{fusion} \times k_{2c,c,i,j} \tag{6}$$

$$G_{b,n,i,j} = \frac{1}{1 + e^{-M_{b,n,i,j}}} \tag{7}$$

where b denotes the batch size, n denotes the number of output channels. We let the coefficient matrix of infrared images $G_{ir} = G$ and the coefficient matrix of visible light images $G_{vi} = 1 - G$, where G_{ir} and G_{vi} denote the contribution level of infrared images and visible light images respectively to predict depth. The coefficient matrix is different under different conditions. For example, $G_{b,i,j,n}^{ir}$ maybe bigger than $G_{b,i,j,n}^{vi}$ in low light conditions while and $G_{b,i,j,n}^{vi}$ maybe bigger than $G_{b,i,j,n}^{ir}$ in the daytime. G_{ir} and G_{vi} are used to represent the contributions of features of infrared images and visible light images as follows:

$$f'_{ir} = f_{ir} \odot G_{ir} \tag{8}$$

$$f'_{vi} = f_{vi} \odot G_{vi} \tag{9}$$

where $\odot$ denotes dot product. The full fused features can be obtained as follows:

$$f'_{fusion} = f'_{ir} + f'_{vi} \tag{10}$$

The full-feature-fusion subnetwork decides the extent we can rely on the infrared features and visible light features respectively to predict depth.

3.3. High-resolution reconstruction subnetwork

Through utilizing the common-feature-fusion subnetwork and the full-feature-fusion subnetwork, we obtain fuses features with the resolution of one-sixteenth of the ground-truth depth map. In order to predict the depth map, we need to map the low-resolution depth map to the corresponding high-resolution depth map with the same resolution as the ground-truth depth map. Checkboard artifacts would emerge and the detail information might be lost if traditional deconvolution is utilized for depth prediction. Inspired by [28,29], we design a high-resolution reconstruction subnetwork based on residual dense convolution blocks. Fig. 4 shows the diagram of the high-resolution reconstruction subnetwork.

Let f'_{fusion} be the low-resolution fused features obtain the full-feature-fusion subnetwork. First, we use a convolutional layer to learn the information from the low-resolution fused features of infrared and visible light images.

$$F_0 = g(f_{low}) \tag{11}$$

where g(·) denotes a composite function of a 3 × 3 convolution and batch normalization (BN). Then F_0 is used as input to residual dense convolution blocks. In residual dense convolution blocks, the output of each layer is connected to all subsequent layers of the current residual dense convolution block directly. It also allows connections from the output of the preceding residual dense convolution block to each layer of the current residual dense convolution block.

$$Input_{n,i} = p_{n-1} + \sum_{c=1}^{i-1} p_{n,c} \tag{12}$$

where n denotes the n-th residual dense convolution block, i denotes the i th layer of the current residual dense convolution block. If the output of each residual dense convolution block has N_0 features and each layer of every block has N features, we concatenate the features produced by the preceding block and preceding layers of the current block, resulting in $N_0 + (i-1)N$ feature maps as the input to the i th layer. In this paper, we set that $N_0 = N = 32$ and each block has three composite operations of convolution, batch normalization and rectified linear unit.

Compared with traditional CNNs, the residual dense convolution blocks can take advantage of the hierarchical features of original low-resolution features to enhance the detail features. After getting the concatenation feature maps, we employ a 1 × 1 convolution operation to learn features of all layers adaptively and make the number of channels the same as F_0 for residual learning. Through three residual dense convolution blocks, we can make the information maximum between layers and get feature maps with more detail information. Then the residual learning is used, we add F_0 and the output of the last block to further enhance the prediction ability of our network. Finally, we design an up-sample layer instead of deconvolution, the output of this layer with the size of 256 × 512 is our predicted depth map. We demonstrate the effectiveness of our high-resolution reconstruction in Section 4.

3.4. Implementation details

To implement our method, we use the deep learning framework of TensorFlow, and train our network using NVIDIA GTX 1080Ti with 11 G video memory.

In order to fairly compare different methods and reduce the influence of the parameter setting to the final depth prediction as much as possible, the initial parameter setting of all methods in this manuscript are the same. Specifically, the weights are all initialized by the truncated normal distribution with the standard deviation of $\sigma = 2/(k \times k \times n_{in})$,

$$X \sim f(x) = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\phi\left(\frac{b-\mu}{\sigma}\right) - \phi\left(\frac{a-\mu}{\sigma}\right)} I \ (a \le x \le b) \tag{13}$$

where k is the kernel size, n_{in} is the number of channels of input, μ is the expectation and (a, b) is the truncated range; the learning rates of all the

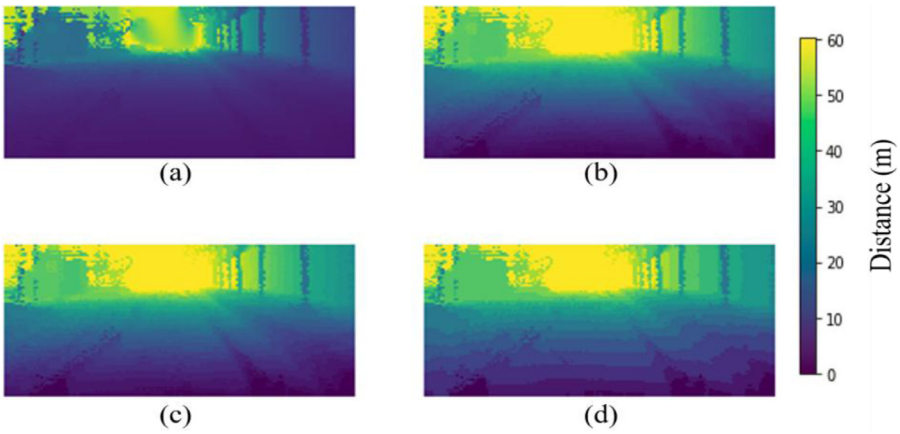

Fig. 5. The ground truth with different levels. (a) original ground truth; (b) 64 hierarchies; (c) 32 hierarchies;(d) 16 hierarchies;.

methods are initialized as $1e-4$ and is reduced by a factor of 0.9 after every 20,000 iterations and the total iterations of all the methods are 80,000. After 80,000 iterations all the methods can converge; the batch size is 4 and use batch normalization after every convolution layer. To prevent overfitting during training, we set the dropout operation after every activation function layer with a ratio of 0.8. For the training loss, cross entropy is selected as our loss and optimizes the network with Adam optimizer.

$$L = \sum_i \left(y_i . \log\left(y_{predicted_i}\right) + \left(1 - y_i\right).\log\left(1 - y_{predicted_i}\right)\right) \tag{14}$$

where L is the loss, $y_{predicted i}$ is the actual prediction and y_i is the expected prediction.

4. Experiments and results

In this section, we introduce our *NUST-SR* dataset and evaluation indicators for the experiments. Through a comprehensive analysis of the proposed IVFuseNet on the dataset, we show the effectiveness of the common-feature-fusion subnetwork, the full-feature-fusion subnetwork, and the high-resolution reconstruction subnetwork. The IVFuseNet achieves best performances on this dataset by comparison with the other nine methods such as fully CNN, multi-scale CNN, etc. [9,10,19,30,33–37].

4.1. Dataset and evaluation

Most of the vision-based unmanned driving depth prediction studies only concerned about the daytime conditions with visible light images. However, the situations of unmanned vehicles driving at low light conditions should also be considered. There are few open-source datasets that contain both infrared and visible images.

In this paper, we use *NUST-SR* dataset, which is composed of the actual road scenes captured while unmanned vehicle driving at daytime and night. The raw *NUST-SR* dataset collected by the vehicle-borne far-infrared camera, the vehicle-borne RGB camera and the Lidar, which contains infrared and visible light images of the resolution of 768×576 and radar data of depth. In the raw dataset, not every point has the depth value, as shown in Fig. 5(a), we filled the points for which there is no depth value first using colorization scheme of Levin et al. in the NYUDepth development kit [9], as shown in Fig. 5(b).

In addition, we use the classification method to predict the depth instead of the regression method as previous works [19,38,39]. The reason is that the real scenes are complex, predicting depth value for each pixel is more difficult than predicting a depth range and our requirements for the accuracy of predicting the depth of distant and near objects are different. The classification method allows us to have higher accuracy for near scenes and lower accuracy for distant scenes [38]. For different numbers of levels, as shown in Fig. 5, the larger the number of levels, the more detailed the ground truth is, and the closer to the original depth map. We can find that the depth map with 16 levels is relatively rough compared to the depth map with 32 or 64 levels, so we focus on comparing the depth map with 32 levels and 64 levels. The depth map with 64 levels is closer to the original ground truth than the depth map with 32 levels, but to a lesser extent, and the number of parameters and FLOPs obtained by training our method with depth map divided into 64 levels are 7% and 22% more than the depth map divided into 32 levels respectively, but the difference between predicted depth maps of the two methods is very small. Therefore, we classify the depth map into 32 levels in the logarithmic space as training labels in this paper, as shown in Fig. 6(c). Finally, we crop the infrared images, visible light images and depth maps to the resolution of 256×512. By preprocessing the raw dataset, we get the *NUST-SR* dataset, which has 6529 registered infrared images, visible light images and the corresponding ground-truth depth maps in the daytime and 5612 at night, respectively.

We evaluate the proposed IVFuseNet on the test images of *NUST-SR* dataset, using four indicators as previous work:

- Root Mean Squared Error (RMSE): $\sqrt{\frac{1}{N}\sum_i^N \left(p_i^{gt} - p_i\right)^2}$
- Mean Relative Error (Rel): $\frac{1}{N}\sum_i^N \frac{|p_i^{gt} - p_i|}{p_i^{gt}}$
- Mean log10 Error (log10): $\frac{1}{N}\sum_i^N |log_{10}^{p_i^{gt}} - log_{10}^{p_i}|$
- Threshold Accuracy: $\max(p_i^{gt}/p_i, p_i/p_i^{gt}) = \delta < \text{threshold}$

In each indicator, p_i^{gt} is the ground-truth depth, p_i denotes the predicted depth and T denotes the total number of pixels. The mean relative error, root mean squared error and mean log_{10} error are all utilized to evaluate the difference between the predicted depth and the corresponding ground-truth depth map, smaller values are better on these evaluation indicators. The threshold accuracy measures the similarity between the predicted depth and the corresponding ground-truth depth map, higher values are better.

4.2. Result for fusion of infrared and visible light image

In order to highlight the effectiveness of the fusion of infrared and visible light images, we make three comparative experiments. Infrared images, visible light images and the fusion of them are used as input to predict depth respectively. For a fair comparison, we replace residual dense convolution blocks with traditional convolution with the same numbers of layers of them to remove impact the of them. And compare the performance of depth predicting only based on visible light images by different methods.

Fig. 7 shows the qualitative result of the above comparative experiments. In the second column, the fusion-based result shows the sharpest edges of the person and the car than the results obtained only based on

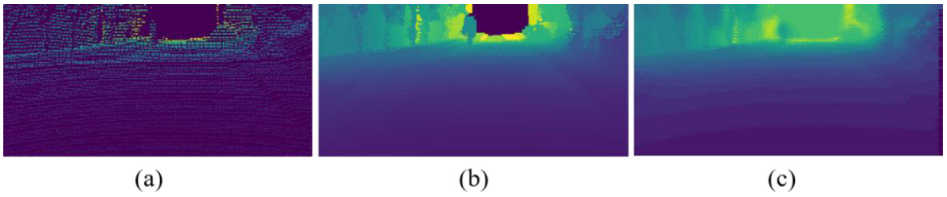

Fig. 6. (a) is the raw depth map generated from Lidar data. (b) is the depth map after filling the points for which there is no depth value in (a). (c) is the depth map classified into 32 levels.

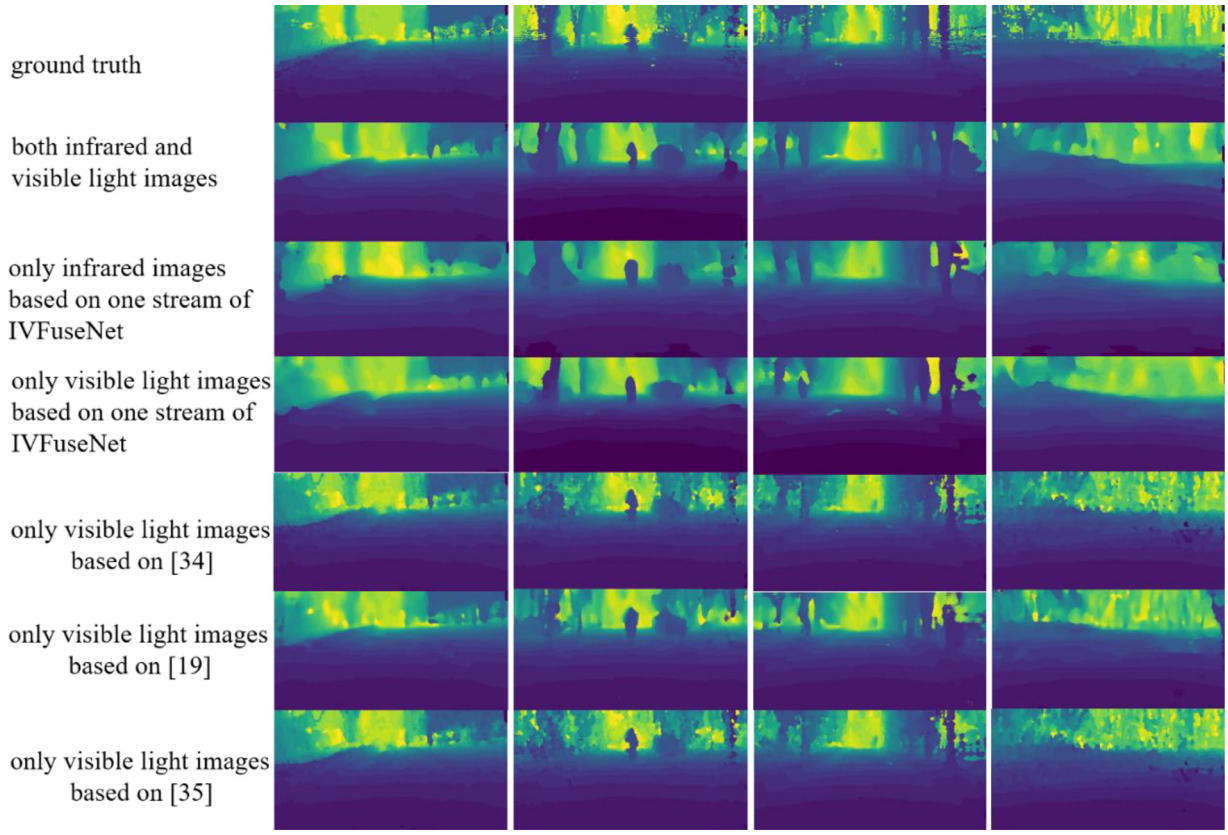

Fig. 7. Comparison of predicting the depth of different scenes based on the fusion images and predicting depth only based on infrared or visible light images.

a single sensor or other methods based on visible light images; In the third column, the predicted trees of the fusion-based result are closest to the ground-truth depth.

Fig. 8 illustrates the performances of testing on the test set of the *NUST-SR* dataset with four indicators. Under the conditions without residual dense convolution blocks, the threshold accuracy is highest and the RMSE, Rel, and log10 error are lowest when we predict depth based on the fusion of infrared and visible light images (the blue curves). This is the result of the combined action of the common-feature-fusion subnetwork and the full-feature-fusion subnetwork.

In the common-feature-fusion subnetwork, the coarse grid search on the test set is used to choose the best configuration of coupled ratios from 0 to 1 with the step size of 0.25. Inspired by [32], we first increase the coupled ratio of deeper layers, if the error rises, we back to the last status, and instead increase the coupled ratio of its immediate lower layer. Note $R_i = 0$ $(i = 1, 2, 3, 4)$ means the two streams are uncorrelated parts, and $R_i = 1$ $(i = 1, 2, 3, 4)$ means the two streams are fully-coupled. As can be seen from Table 2, k_i $(i = 1, 2, 3, 4)$ means the number of partial coupled filters. When we set the coupled ratios as $R_1 = 0$, $R_2 = 0.25$, $R_3 = 0.5$, $R_4 = 0.75$, the Rel is the lowest. The result shows that the similarity between infrared images features and visible light features is stronger with the increasing of convolutional layers.

Simultaneously, comparing with the uncoupled situation in the first row of Table 2, the Rel decrease 3.5%. It demonstrates that infrared and visible light images as the "auxiliary variables" of each other enhance the features of each other to obtain more sufficient and effective information and decrease the predicted error.

In the full-feature-fusion subnetwork, we adopt the adaptive weighted fusion strategy to fuse the full features of infrared images and visible light images extracted by the common-feature-fusion subnetwork. In order to demonstrate that the method can tackle the problem of different contributions of infrared and visible light images for depth prediction in different conditions, we design two groups of comparative experiments.

Fig. 9 shows the visualization of convolutional layers. Fig. 8(c) and (d) are feature maps obtained by the last convolutional layers before the fusion operation, and Fig. 8(e) are fusion features of infrared and visible light images. We choose the tenth channel to visualize. In the first row, the scene is in the daytime. It can be seen that the edge of the road and the trees are extracted by the convolutional layer, the features of visible images are clearer than that of infrared images. By observing the fusion features, it can be easily known that the features of visible light images make more contributions. In the second row, the scene is in the night. Obviously, the visible light images of scenes in the night have few available features, and most of the contributions are made by infrared images. This demonstrates that our adaptive weighted fusion method considers the different contributions of the two kinds of images on delineating different categories of objects in different scenes. Whether it is

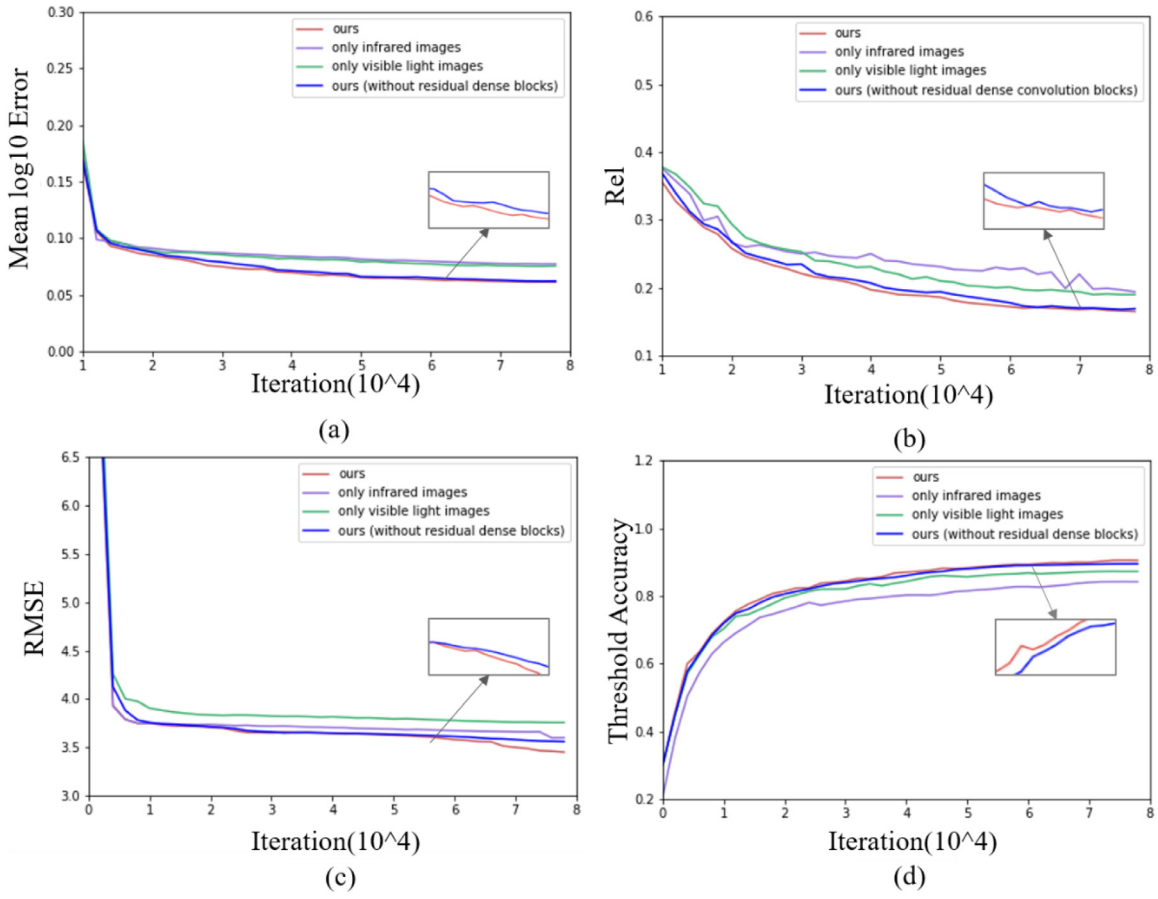

Fig. 8. The comparison of accuracy and error under different input.

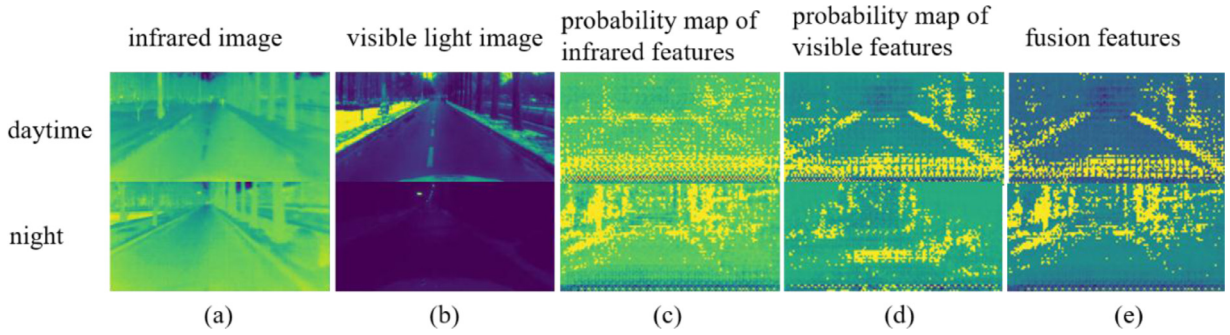

Fig. 9. Visualization of convolutional layers.

in the daytime or night, the fusion features of two kinds of images are clearer and have less noise.

Many researchers use the methods based deep learning [22,24,40] or traditional machine learning [23,44,45] to fuse two different kinds of images. In Fig. 10, we compare these methods to our fusion method. By comparison, we find that the predicted depth maps based on our fusion method have considerable advantages in the details and the edges of different objects.

Interestingly, in the fourth column, the predicted depth map is better than the corresponding ground-truth depth. Because not every point of scatter depth maps has the depth value. Although we filled the points for which there is no depth value, the values of these points are not right completely, objects are not depicted properly. The infrared and visible light images are more precise for the shape of objects, we can extract the detailed feature of the person with the training of IVFuseNet. Therefore, the predicted depth map is better than the corresponding ground-truth depth can be understood.

4.3. Result for residual dense convolution blocks

To demonstrate the effectiveness of residual dense convolution blocks in the high-resolution reconstruction subnetwork, we compare the performance of IVFuseNet and the network which replace the residual dense convolution blocks with traditional convolutional layers in Section 4.2.

By the comparison, it can be seen that the IVFuseNet with residual dense convolution blocks has large improvement. As shown in Fig. 11, when we use residual dense convolution blocks, the contour details of small objects are shaper, such as the person and the car in the first column and the people in the fourth column. Conversely, when we predict

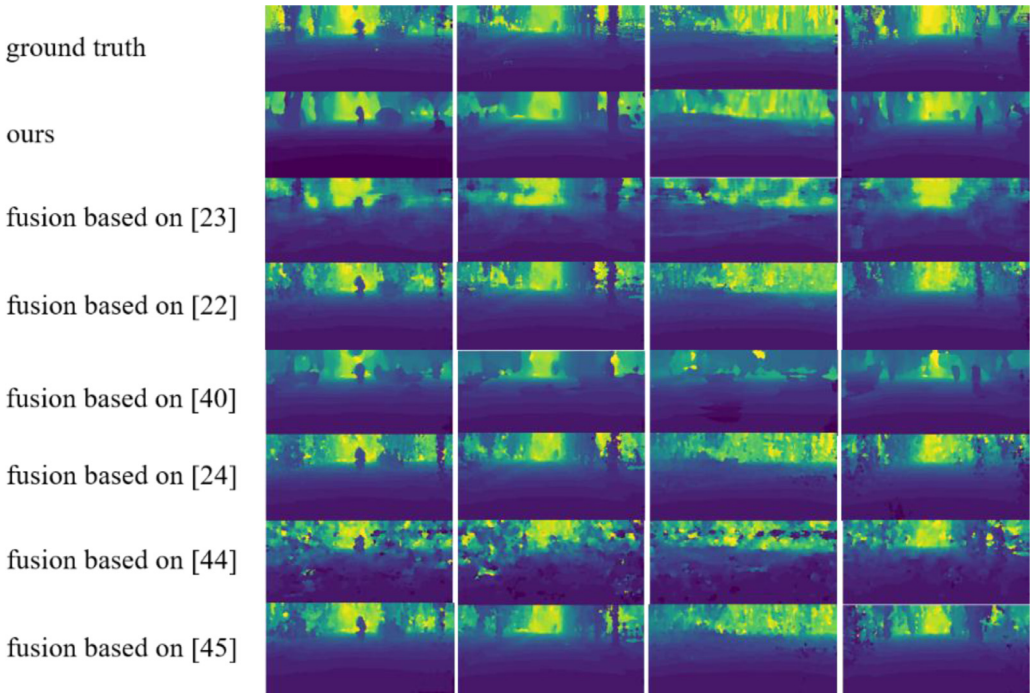

Fig. 10. Comparison of different fusion methods in different scenes.

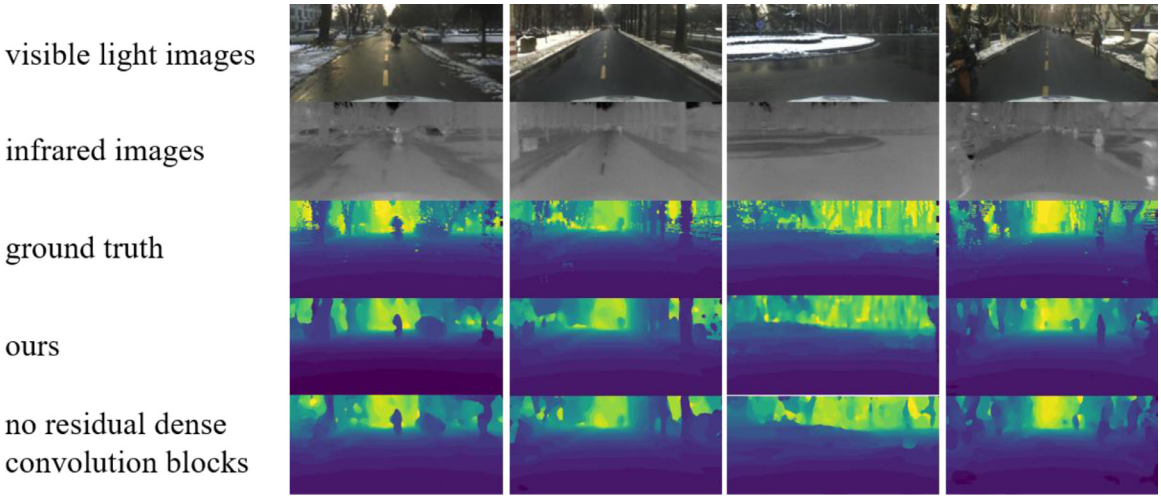

Fig. 11. Compare our method (IVFuseNet) with the method which removes of residual dense convolutional blocks.

depth with traditional convolution, the trees in the second column are thinner than these in the ground-truth depth map and the shape is distorted. In Fig. 7, the curves of the four indicators show that the residual dense convolution blocks further decrease the error and increase the threshold accuracy based on the fusion (the red curves). This demonstrates that the residual dense convolutional blocks can make full use of the hierarchical features of original low-resolution features to help recover the finer details.

4.4. Comparison with related methods

In Fig. 12, we qualitatively compare the performance of depth prediction of our proposed method (IVFuseNet) with that of the other methods. We can clearly see that the improvement of the edge quality in the predicted depth maps. Whether the trees in the second and third column, the car in the first column or the person in the fourth column of depth maps, our method has clearer edge than the other methods. Through fusing the information of infrared images and visible light images, IVFuseNet can increase the detail information of infrared images and reduce the redundant features of visible light images. Furthermore, the residual dense convolution blocks enhance the fusion features and recover the details.

To more scientifically illustrate the performance of our method, we also quantitatively compare the ten methods in Table 3 using the evaluate indicators reported in Section 4.1. It can be observed that IVFuseNet achieves the best result on all metrics, especially in threshold accuracy and RMSE indicators.

In order to analysis the computational complexity and time consumption we compare the parameters and the floating points operations (FLOPs) of the proposed method (IVFuseNet) and the compared methods in Table 4. We find that the IVFuseNet has the similar parameters and FLOPs with [9] and [10]. Although the FLOPs of [30] is least, the performance of the evaluate indicators are not very well. And it can be observed that if we replace residual dense convolution blocks with

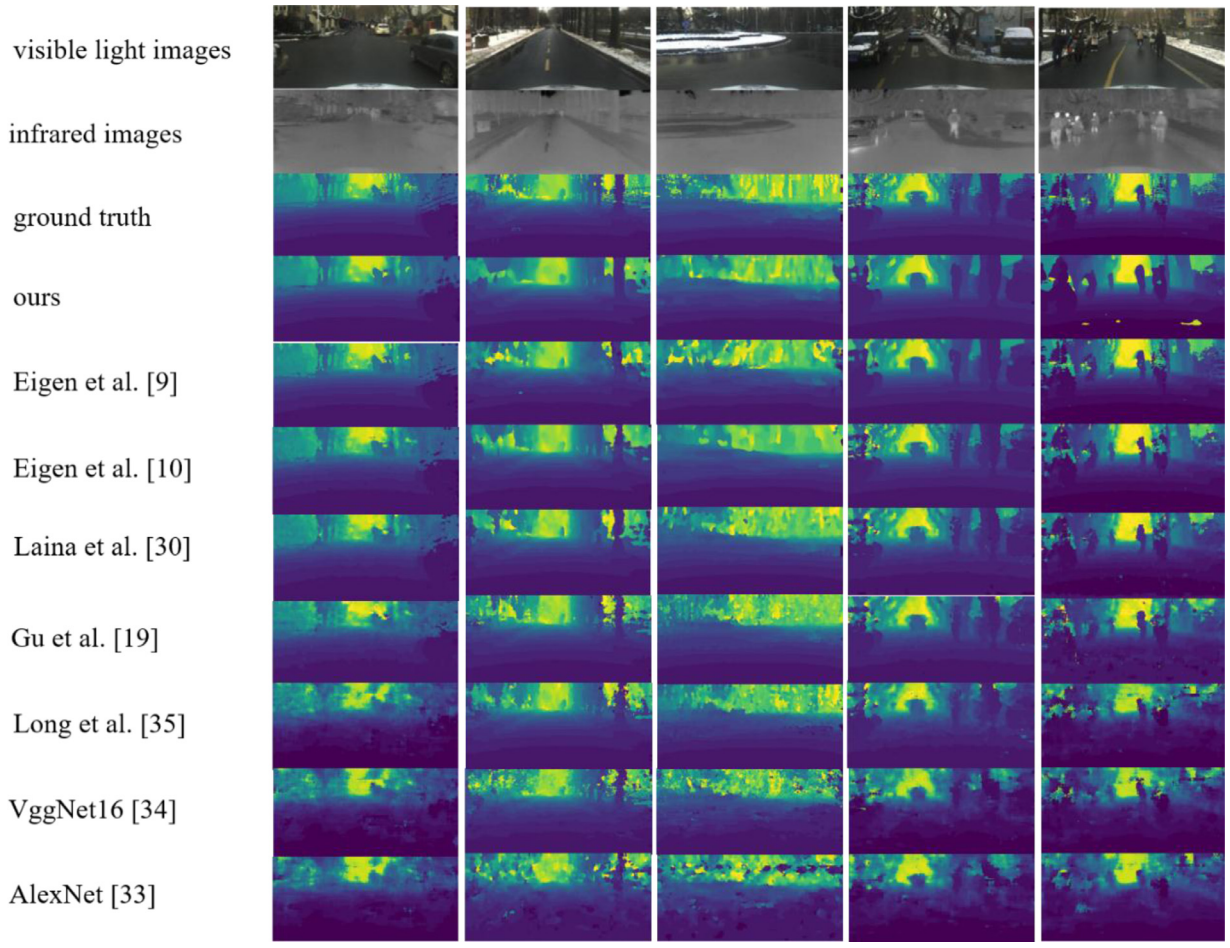

Fig. 12. Depth prediction on *NUST-SR* dataset. Qualitative results showing predictions using eight methods.

Table 3
Comparison with the related seven methods using four evaluate indicators.

	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$	Rel	log10	RMSE
Eigen et al. [9]	0.7590	0.8895	0.9306	0.1935	0.0690	3.7376
Eigen et al. [10]	0.6166	0.7926	0.8640	0.1918	0.0672	3.6900
Fu et al. [36]	0.7629	0.8935	0.9317	0.1853	0.0676	3.5369
Laina et al. [30]	0.6757	0.8507	0.9044	0.2550	0.0902	4.3087
Gu et al. [19]	0.6358	0.8092	0.8795	0.1836	0.0627	3.5660
Long et al. [35]	0.6345	0.8076	0.8741	0.1724	0.0613	3.5530
Wang et al. [37]	0.6352	0.8055	0.8795	0.1903	0.0645	3.5672
VggNet16 [34]	0.6360	0.8126	0.8815	0.1919	0.0679	3.5704
AlexNet [33]	0.5056	0.7437	0.8343	0.3301	0.1192	4.8526
Ours	**0.7898**	**0.9048**	**0.9386**	**0.1651**	**0.0613**	**3.4513**
	higher	is	better	lower	is	better

Table 4
Comparison of the parameters and GFLOPs between IVFuseNet and other seven related methods.

Method	Ours (only fusion part)	Ours	Eigen et al. [9]	Eigen et al. [10]	Fu et al. [36]
Parms	7.02E+06	7.33E+06	6.71E+06	6.61E+06	9.53E+06
GFLOPs	4.27	25.31	23.15	21.48	67.09
Method	Laina et al. [30]	Gu et al. [19]	Long et al. [35]	Wang et al. [37]	VggNet16 [34]
Parms	8.54E+06	4.96E+07	6.86E+07	4.88E+07	2.67E+07
GFLOPs	10.06	286.18	993.16	87.65	987.86

traditional convolution with the same numbers of layers of them, the IVFuseNet contain least parameters and the FLOPs of IVFuseNet is less 57% than the FLOPs of [30]. It illustrates that the filters coupling between the infrared stream and visible light stream in the common-feature-fusion subnetwork reduces the parameters and computational complexity significantly. The residual dense convolution blocks only enhance the detail information to some extent. Therefore, we can only use IVFuseNet without residual dense convolution blocks when we want to reduce the complexity and do not have enough memory space.

5. Conclusion

In this paper, we have proposed a CNN-based architecture called IVFuseNet, which can fuse the complementarity of infrared and visible light images adaptively to address the depth prediction problem under various light conditions. The fusion method we proposed mainly includes two aspects. First, we design partially coupled filters in the common-feature-fusion subnetwork, features of infrared and visible light images are enhanced by the "auxiliary variable". Second, in order to consider the different contributions of infrared and visible light images in different light conditions, we design the adaptive weighted fusion method. Moreover, we introduce three residual dense convolution blocks to further recover the details and increase the accuracy of depth prediction. The effectiveness of IVFuseNet has been demonstrated on our *NUST-SR* dataset. Compared with other methods, we obtain the best performance on this dataset.

Further works will be contributed to the collection of new datasets and the design novel architectures for depth prediction. Nowadays the depth prediction for unmanned driving is mainly focused on daytime applications and the datasets available are designed under very good light conditions. New datasets under different weather conditions and light conditions should be developed for further research. In this paper, the sequential information of the images is not considered. In the future, we will try to adopt recurrent neural networks to deal with the correlation information between different frames of the image sequence.

Declaration of Competing Interest

The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.

Acknowledgements

This research is sponsored by National Natural Science Foundation of China (61375007 and 61375012) and Basic Research Programs of Science and Technology Commission Foundation of Shanghai (15JC1400600).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.inffus.2019.12.014.

References

[1] K. Maham, S. Hassan, A. Syed Irfan, J. Iqbal, Stereovision-based real-time obstacle detection scheme for unmanned ground vehicle with steering wheel drive mechanism, in: 2017 International Conference on Communication, Computing and Digital Systems (C-CODE), 2017, pp. 380–385, doi:10.1109/C-CODE.2017.7918961.

[2] J. Woo, N. Kim, Vision based obstacle detection and collision risk estimation of an unmanned surface vehicle, in: 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), 2016, pp. 461–465, doi:10.1109/URAI.2016.7734083.

[3] X. Han, J. Lu, Y. Tai, C. Zhao, A real-time LIDAR and vision based pedestrian detection system for unmanned ground vehicles, in: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015, pp. 635–639, doi:10.1109/ACPR.2015.7486580.

[4] U. Rasheed, M. Ahmed, M.J. Afridi, F. Kunwar, Road trajectory mining and autonomous steering control for vision-based unmanned vehicles, in: 2010 10th International Conference on Intelligent Systems Design and Applications, 2010, pp. 197–202, doi:10.1109/ISDA.2010.5687267.

[5] C. Chuang, J. Li, Applying grey-fuzzy-fuzzy rule to machine vision based unmanned automatic vehicles, 2010. doi:10.1007/978-3-642-16584-9_25.

[6] S. Zhang, X. Wang, Human detection and object tracking based on histograms of oriented gradients, in: 2013 Ninth International Conference on Natural Computation (ICNC), 2013, pp. 1349–1353, doi:10.1109/ICNC.2013.6818189.

[7] K. Lee, C. Choo, H. See, Z. Tan, Y. Lee, Human detection using histogram of oriented gradients and human body ratio estimation, in: 2010 3rd International Conference on Computer Science and Information Technology, 2010, pp. 18–22, doi:10.1109/ICCSIT.2010.5564984.

[8] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), 1, 2005, pp. 886–893, doi:10.1109/CVPR.2005.177.

[9] D. Eigen, C. Puhrsch, R. Fergus, Depth map prediction from a single image using a multi-scale deep network, in: NIPS'14 Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. 2, 2014, pp. 2366–2374.

[10] D. Eigen, R. Fergus, Predicting Depth, Surface normals and semantic labels with a common multi-scale convolutional architecture, in: 2015 IEEE International Conference on Computer Vision (ICCV), 2015, pp. 2650–2658, doi:10.1109/ICCV.2015.304.

[11] J. Li, R. Klein, A. Yao, A two-streamed network for estimating fine-scaled depth maps from single RGB images, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 3392–3400, doi:10.1109/ICCV.2017.365.

[12] Y. Kuznietsov, J. Stückler, B. Leibe, Semi-Supervised deep learning for monocular depth map prediction, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2215–2223, doi:10.1109/CVPR.2017.238.

[13] C. Godard, O. Mac Aodha, G.J. Brostow, Unsupervised monocular depth estimation with left-right consistency, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 6602–6611, doi:10.1109/CVPR.2017.699.

[14] J. Ma, J. Zhao, Y. Ma, Non-rigid visible and infrared face registration via regularized Gaussian fields criterion, Pattern Recognit. 48 (2015) 772–784.

[15] J. Ma, C. Chen, C. Li, Infrared and visible image fusion via gradient transfer and total variation minimization, Inf. Fusion 31 (2016) 100–109.

[16] L. Xi, S. Sun, L. Li, F. Zou, Depth estimation from monocular infrared images based on SVM model, Laser Infrared 42 (2012) 1311–1315.

[17] S. Sun, L. Li, L. Xi, Depth estimation from monocular infrared images based on BP neural network model, Comput. Vision Remote Sens. (CVRS) (2012) 237–241, doi:10.1109/CVRS.2012.6421267.

[18] S. Sun, L. Li, H. Zhao, Depth estimation from monocular vehicle infrared images based on KPCA and BP neural network, Infrared Laser Eng. 49 (2013) 2348–2352.

[19] T. Gu, H. Zhao, S. Sun, Depth estimation of an infrared image based on interframe information extraction, Laser Optoelectron. Progr. 55 (2018) 169–178.

[20] A.V. V.anmali, V.M. Gadre, Visible and NIR image fusion using weight-map-guided Laplacian-Gaussian pyramid for improving scene visibility, Sādhanā 42 (2017) 1063–1082, doi:10.1007/s12046-017-0673-1.

[21] X. Han, L. Zhang, L. Du, Fusion of infrared and visible images based on discrete wavelet transform, Sel. Pap. Photoelectron. Technol. Committee Conf. (2015), doi:10.1117/12.2216054.

[22] H. Li, X. Wu, DenseFuse: a fusion approach to infrared and visible images, IEEE Trans. Image Process. (2019) In press, doi:10.1109/TIP.2018.2887342.

[23] Y. Liu, Z. Wang, Simultaneous image fusion and denoising with adaptive sparse representation, Image Process. Lett. 9 (2014) 347–357, doi:10.1049/iet-ipr.2014.0311.

[24] K.R. Prabhakar, V.S. Srikar, R.V. Babu, DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 4724–4732, doi:10.1109/ICCV.2017.505.

[25] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: a survey, Inf. Fusion 45 (2019) 153–178.

[26] Z. Zhou, B. Wang, S. Li, M. Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters, Inf. Fusion 30 (2016) 15–26.

[27] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: a generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.

[28] G. Huang, Z. Liu, K. Weinberger, Densely connected convolutional networks, in: IEEE Conference in Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4700–4708, doi:10.1109/CVPR.2017.243.

[29] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, Y. Fu, Residual dense network for image super-resolution, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2472–2481, doi:10.1109/CVPR.2018.00262.

[30] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, N. Navab, Deeper depth prediction with fully convolutional residual networks, in: 2016 Fourth International Conference on 3D Vision (3DV), 2016, pp. 239–248, doi:10.1109/3DV.2016.32.

[31] C. Li, J. Hu, Y. Han, Infrared and visible light images fusion utilizing object detection, J. Image Signal Process. 4 (2015) 47–52, doi:10.12677/JISP.2015.43006.

[32] Z. Wang, S. Chang, Y. Yang, D. Liu, T.S. Huang, Studying very low resolution recognition using deep networks, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4792–4800, doi:10.1109/CVPR.2016.518.

[33] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, Conf. Workshop Neural Inf. Process. Syst. (2012) 5, doi:10.1145/3065386.

[34] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, Comput. Sci. (2014).

[35] D.T. Long J, E. Shelhamer, Fully convolutional networks for semantic segmentation, IEEE Trans. Pattern Anal. Mach. Intell. (2015) 3431–3440, doi:10.1109/CVPR.2015.7298965.
[36] H. Fu, M. Gong, C. Wang, K. Batmanghelich, D. Tao, Deep ordinal regression network for monocular depth estimation, in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 2002–2011, doi:10.1109/CVPR.2018.00214.
[37] P. Wang, X. Shen, S. Cohen, B. Price, A.L. Yuille, Towards unified depth and semantic prediction from a single image, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 2800–2809, doi:10.1109/CVPR.2015.7298897.
[38] S. Wu, H. Zhao, S. Sun, Depth estimation from infrared video using local-feature-flow neural network, Int. J. Mach. Learn. Cybern. (2018) 1–10. doi: 10.1007/s13042-018-0891-9.
[39] S. Wu, H. Zhao, S. Sun, Depth estimation from monocular infrared video based on bi-recursive convolutional neural network, Acta Optiva Sinica 37 (2017) 254–262.
[40] J. Chen, J. Wu, J. Konrad, P. Ishwar, Semi-Coupled two-stream fusion convnets for action recognition at extremely low resolutions, in: IEEE Winter Conference on Applications of Computer Vision (WACV), 2017, pp. 139–147, doi:10.1109/WACV.2017.23.
[41] D. Xu, E. Ricci, W. Ouyang, X. Wang, N. Sebe, Multi-Scale continuous CRFs as sequential deep networks for monocular depth estimation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 161–169, doi:10.1109/CVPR.2017.25.
[42] B. Li, C. Shen, Y. Dai, M. He, Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 1119–1127, doi:10.1109/CVPR.2015.7298715.
[43] Y. Cheng, R. Cai, Z. Li, Locality-Sensitive deconvolution networks with gated fusion for RGB-d indoor semantic segmentation, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1475–1483.
[44] H. Li H, X. Wu, Multi-focus Image Fusion using dictionary learning and Low-Rank Representation, 2018.
[45] H. Li, X. Wu, Multi-focus Noisy Image Fusion using Low-Rank Representation, 2018.