Wanpeng XU, Ling ZOU, Lingda WU, Yue QI, Zhaoyong QIAN, 2022. Depth estimation using an improved stereo network. *Frontiers of Information Technology & Electronic Engineering*, 23(5):777-789. https://doi.org/10.1631/FITEE.2000676

# Depth estimation using an improved stereo network

Front Inform Technol Electron Eng

**Key words:** Monocular depth estimation; Self-supervised; Image reconstruction

Corresponding author: Ling ZOU
E-mail: zouling@bfa.edu.cn

# Motivation

- ResNet, which serves as a backbone network, has some structural deficiencies when applied to downstream fields, because its original purpose was to cope with classification problems.
- No pixels that violate camera motion assumptions are traditionally expected when using stereo pairs as the input for self-supervised methods. However, in low-texture areas, performance may be deteriorated.

Front Inform Technol Electron Eng

# Contribution

- An improved network architecture that performs self-supervised monocular depth estimation using stereo pairs is proposed to better propagate information through the network's layers.
- A novel disparity consistency loss ignores the training pixels of images in the low-texture area.
- A novel input method that inputs left and right images randomly expands the dataset.

Front Inform Technol Electron Eng

# Framework

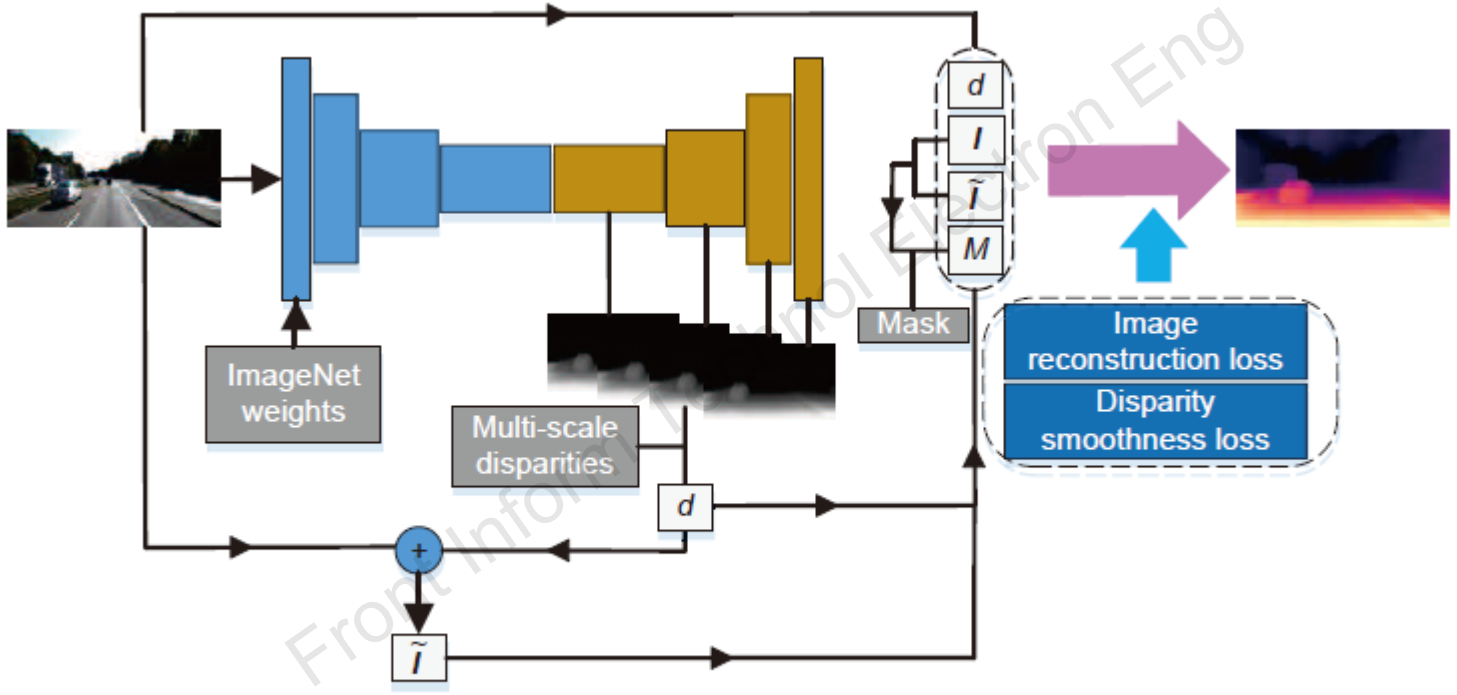
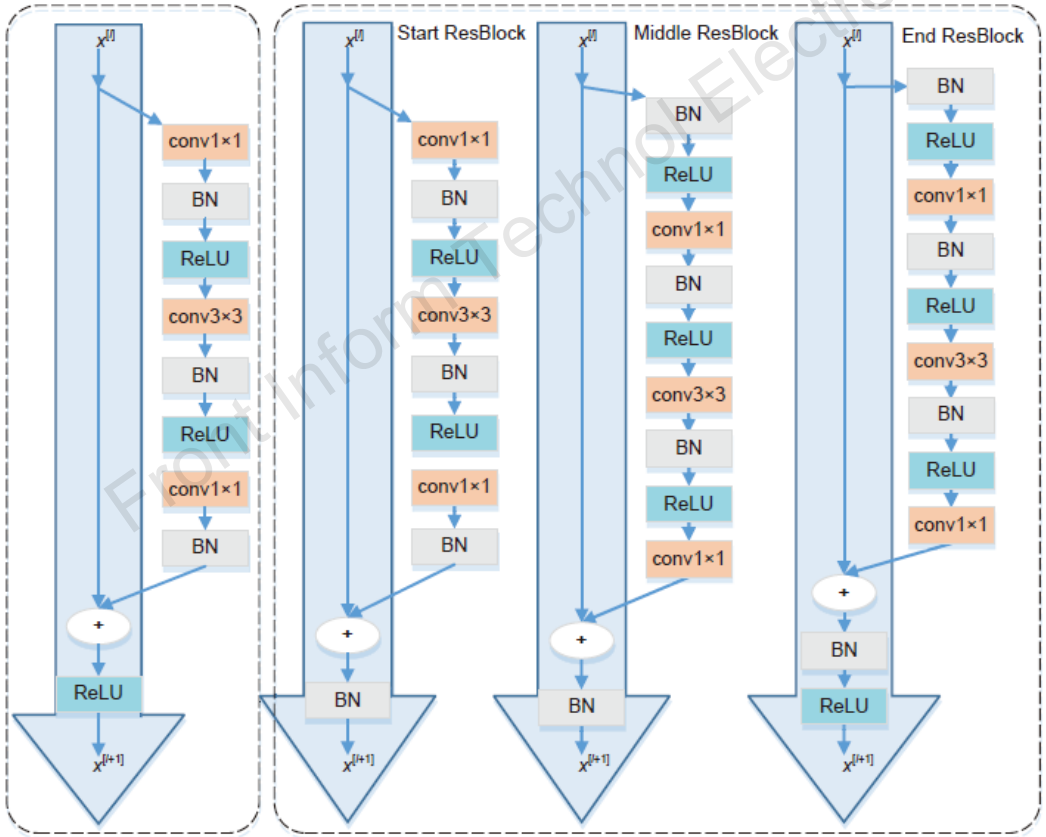



Fig. 2 Illustration of our deep estimation model architecture

# Method

- Improved information flow through the network

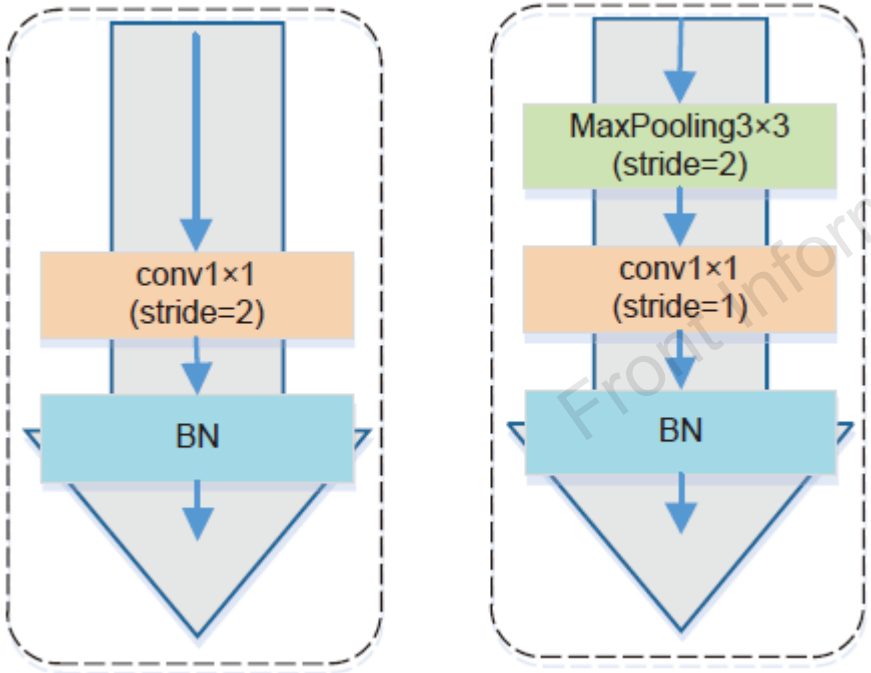
  In our method, only four ReLUs on the main propagation path can retain as many effective backpropagation nodes as possible while maintaining the model's learning ability.



# Method (Cont'd)

- Improved projection shortcut

In the projection shortcut, we use a meaningful way to select the feature maps that go to the next stage.




- Grouped building block

The conv3×3 acquires the maximum number of channels and the grouped convolution is adopted to perform the convolution operation independently for each group.




Front Inform Technol Electron Eng

# Method (Cont'd)

- Training loss

  We use a binary mask to filter out the pixels whose appearance does not change between the stereo pairs. This method can eliminate the effect of pixels in a low-texture region in the network. So, the loss can be written in the following form:

$$L_s = L_{\mathrm{p}} \odot M_{\mathrm{p}} + \lambda L_{\mathrm{smooth}},$$

$$L_{\mathrm{p}} = \frac{1}{N} \sum \left( \alpha \frac{1 - \mathrm{SSIM}(\boldsymbol{I}, \hat{\boldsymbol{I}})}{2} + (1 - \alpha) \| \boldsymbol{I} - \hat{\boldsymbol{I}} \| \right),$$

$$M = L_{\mathrm{p}}(\boldsymbol{I}, \tilde{\boldsymbol{I}}) < L_{\mathrm{p}}(\boldsymbol{I}^{\mathrm{l}}, \boldsymbol{I}^{\mathrm{r}}),$$

$$L_{\mathrm{smooth}} = \frac{1}{N} \sum \left( |\partial_x d|^{-|\partial_x I|} + |\partial_y d|^{-|\partial_y I|} \right),$$

Front Inform Technol Electron Eng

where $L_{\mathrm{p}}$ is the appearance matching loss, $L_{\mathrm{smooth}}$ is the smooth loss, and $M$ is the mask.

# Major results

**Table 1 Comparison of performances reported on the KITTI dataset**

| Method | Training | AbsRel | SqRel | RMSE | $RMSE_{log}$ | $\delta_1 < 1.25$ | $\delta_2 < 1.25^2$ | $\delta_3 < 1.25^3$ |
|---|---|---|---|---|---|---|---|---|
| Garg | S | 0.152 | 1.226 | 5.849 | 0.246 | 0.784 | 0.921 | 0.967 |
| StrAT | S | 0.128 | 1.019 | 5.403 | 0.227 | 0.827 | 0.935 | 0.971 |
| Monodepth | S | 0.128 | 1.038 | 5.355 | 0.223 | 0.833 | 0.939 | 0.972 |
| 3Net | S | 0.119 | 1.201 | 5.888 | 0.208 | 0.844 | 0.941 | **0.978** |
| MonoResMatch | S | 0.116 | 0.986 | 5.098 | 0.214 | 0.847 | 0.939 | 0.972 |
| SuperDepth | S | 0.112 | 0.875 | 4.958 | 0.207 | 0.852 | 0.947 | 0.977 |
| Monodepth2 | S | 0.106 | 0.854 | 4.835 | 0.203 | 0.873 | 0.950 | 0.976 |
| RefineDistill | S | 0.098 | 0.831 | 4.656 | 0.202 | 0.882 | 0.948 | 0.973 |
| Ours (IB) | S | 0.102 | 0.794 | 4.710 | 0.200 | 0.877 | 0.953 | 0.977 |
| Ours (IGC) | S | 0.104 | 0.829 | 4.800 | 0.202 | 0.875 | 0.952 | 0.976 |
| Ours (IB+IGC) | S | 0.102 | 0.790 | 4.684 | 0.198 | 0.878 | 0.954 | 0.977 |
| Ours (IB+IGC) HR | S | **0.097** | **0.732** | **4.519** | **0.194** | **0.884** | **0.956** | **0.978** |
| SfMLearner | M | 0.183 | 1.595 | 6.709 | 0.270 | 0.734 | 0.902 | 0.959 |
| GeoNet | M | 0.149 | 1.060 | 5.567 | 0.226 | 0.796 | 0.935 | 0.975 |
| DF-Net | M | 0.150 | 1.124 | 5.507 | 0.223 | 0.806 | 0.933 | 0.973 |
| EPC++ | M | 0.141 | 1.029 | 5.350 | 0.216 | 0.816 | 0.941 | 0.976 |
| Struct2depth | M | 0.141 | 1.026 | 5.291 | 0.215 | 0.816 | 0.945 | 0.979 |
| Monodepth2 | M | 0.115 | 0.903 | 4.863 | 0.193 | 0.877 | 0.959 | 0.981 |
| PackNet-SfM | M | 0.111 | 0.785 | 4.601 | 0.189 | 0.878 | 0.960 | 0.982 |
| PackNet-SfM HR | M | 0.107 | 0.803 | 4.566 | 0.197 | 0.876 | 0.957 | 0.979 |
| MonoResMatch | SGM+S | 0.111 | 0.867 | 4.714 | 0.199 | 0.864 | 0.954 | 0.979 |
| EPC++ | MS | 0.128 | 0.935 | 5.011 | 0.209 | 0.831 | 0.945 | 0.979 |
| Monodepth2 HR | MS | 0.106 | 0.806 | 4.630 | 0.193 | 0.876 | 0.958 | 0.980 |
| DepthHints | SGM+MS | 0.105 | 0.769 | 4.627 | 0.189 | 0.875 | 0.959 | 0.982 |
| DepthHints HR | SGM+MS | 0.098 | 0.702 | 4.398 | 0.183 | 0.887 | 0.963 | 0.983 |
| Ours (IB+IGC) | MS | 0.101 | 0.723 | 4.463 | 0.180 | 0.900 | 0.965 | 0.983 |
| Ours (IB+IGC) HR | MS | **0.096** | **0.632** | **4.241** | **0.173** | **0.906** | **0.967** | **0.984** |

Front Inform Technol Electron Eng

# Major results (Cont'd)

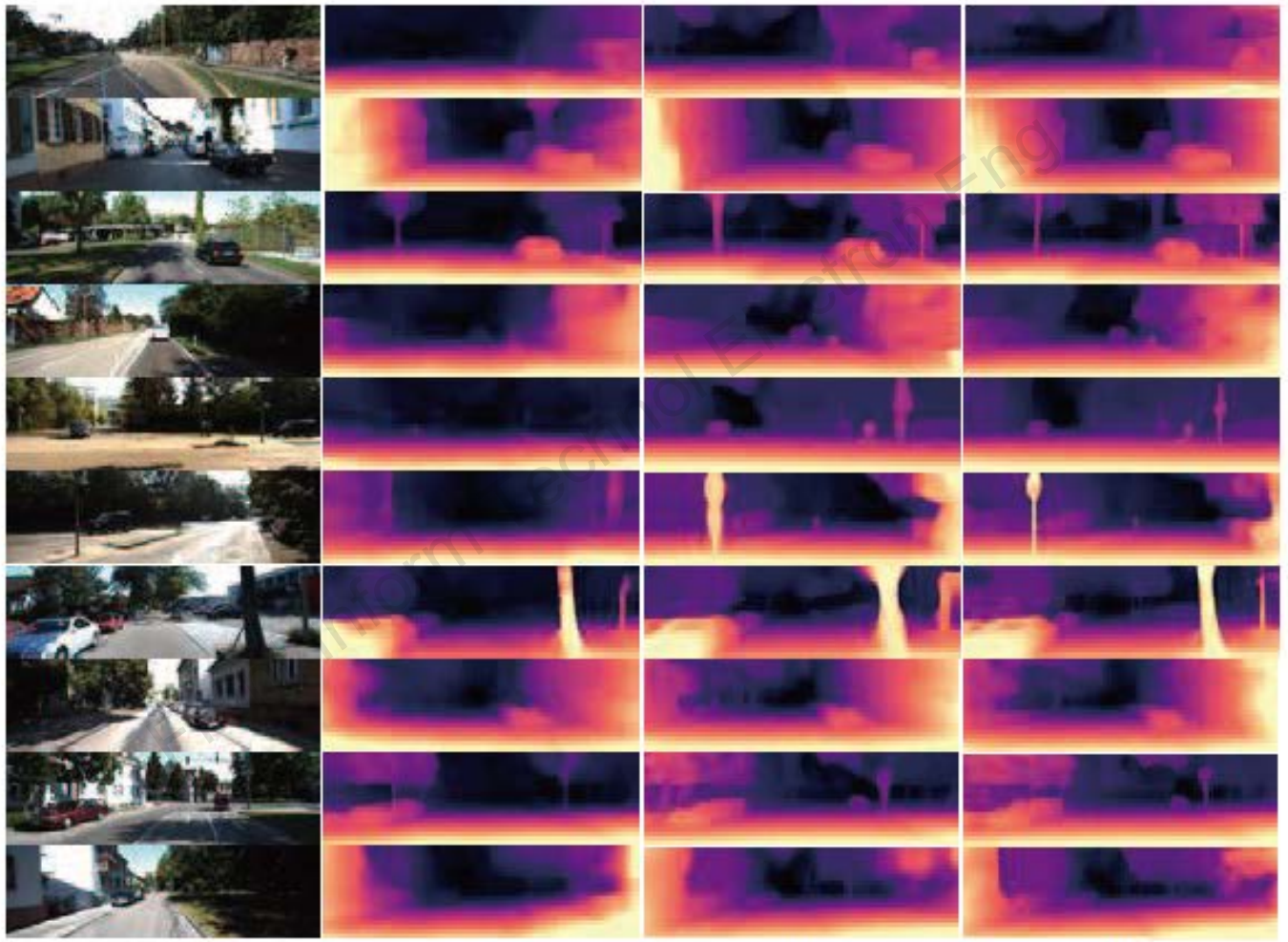



**Fig. 7 Qualitative comparison of our approach with state-of-the-art methods on frames from the KITTI dataset: (a) original images; (b) struct2depth; (c) monodepth2; (d) ours**

Our model produced better performance in low-texture areas, predicted sharper edges, and retained more spatial information

# Conclusions

We adopted a novel monocular depth estimation backbone network, abandoning traditional ResNet based methods. This boosted the propagation of disparate information through network layers and improved the utilization of efficient information. Furthermore, the binary mask used to deal with the pixels in the low-texture area eliminated the interference of invalid disparities in training. Finally, for more extensive and accurate general feature representation, we loaded weights that were pre-trained on ImageNet. By combining the above improvements, we produced state-of-the-art results on the Eigen split of the KITTI driving dataset using stereo pairs in a self-supervised manner.

Front Inform Technol Electron Eng