ISSN: 2664-9640

Depth estimation of monocular endoscopic images based on improved HS-ResNet

Chongchong Zhang, Jinzhao Lin

Chongqing University of Posts and Telecommunications, Chongqing 400000, China

Abstract

In minimally invasive surgery, due to the complex tissue structure of the human body, the blurred texture of internal organs, and the complex gradients, the use of only a 2D display platform will increase the risk of surgery and reduce the performance of computer-assisted algorithms. Therefore, it is necessary to perform three-dimensional endoscopic images. It is shown that this method can increase the doctor's surgical field of vision, and the observation is more three-dimensional. The premise of 3D reconstruction needs to be based on accurate depth information. Therefore, this paper proposes a depth estimation method for monocular endoscope based on improved HS-Net network. Laparoscopic images are used as endoscopic images, real laparoscopic images provided by Hamlyn Center Laparoscopy/Endoscope Video DataSet (http://hamlyn.doc.ic.ac.uk/vision/), data The set contains a total of 32,400 pairs of binocular endoscopic images. A novel internal feature fusion mechanism and effective attention module are applied to the network for depth estimation, and the SSIM has achieved 0.8826 ± 0.0678, which is 40% higher than the basic model. It can be seen in the depth image that more detailed infor

Keywords

Depth estimation; endoscopic image; deep learning.

1. Introduction

At present, most hospitals in China are equipped with endoscopes[1] to help doctors perform minimally invasive surgery on abdominal cavity, thoracic cavity, ear, nose and throat and develop a variety of derivative functions[2-3]. Endoscopy has obvious advantages. Its advantage is that the patient's abdominal cavity and thoracic cavity do not need to be cut with a scalpel. The operation can be carried out only by opening three small holes in the target area of the operation. Due to the fuzzy texture and complex structure of medical image, it needs better and higher resolution information representation. At the same time, the internal structure of human body is relatively fixed and needs low resolution information representation. Therefore, u-net, which can combine high and low resolution information at the same time, is used as the basic framework of this model. Through the encoder network, it obtains low-resolution information after multiple down sampling. After concatenate operation, it can directly transfer high-resolution information from the encoder to the same height decoder. Therefore, u-net[4] is widely used in medical image processing. Taking the u-net structure used in monodepth as the main framework, monodepth realizes binocular vision unsupervised depth estimation, The proposed u-net network is based on the classical VGG model, with simple structure and complete function. It is widely cited by many articles on endoscopic image processing[5].

For extracting features from the input images[6] proposed DispNet which was based on U-Net[7], a typical encoder-decoder architecture. Monodepth2[8] proposed a feature encoder based on ResNet[9] which has since become the standard approach. To increase the robustness of the photometric loss, Shu et al.[10] used an external network to transform a reference frame and

ISSN: 2664-9640

target frames into another domain in which there are better alternative representations for texture-less regions. Guizilini et al.[11] introduced 3D convolutions to construct packing and unpacking blocks, which are the replacement of standard down-sample and up-sample operations and preserve more details in feature maps than those of 2D convolutions. Inspired by these ideas, we investigate a network architecture which has two key attributes: depth estimation and an internal mechanism which evolves multiple chances for feature fusion. Therefore, we choose HS-Net[12] as our new encoder blueprint. HS-Net is able to learn high-resolution representations from images that are both semantically and spatially descriptive, and has been successfully applied to human pose estimation, semantic segmentation and object detection.

2. Network architecture

2.1. Residual encoders

In this paper, ResNet is used as the network convolution layer to construct the u-net structure. The characteristics of the endoscope image are extracted by the encoder, and the size of the image is restored to the original size by the decoder. The encoder first performs preprocessing convolution on the input RGB image, and the convolution kernel size is 7 × 7. The step length is 2 and the zero filling is 3. After preprocessing, the image is normalized in batch, and then 4 times of convolution kernel is 3 × 3. After five times of convolution, the characteristic dimensions of the size of the convolution kernel are 16, 32, 64, 128 and 256 respectively. As shown in the figure, HS-ResNet is composed of multiple segmentation and splicing operations, and the segmentation and splicing operations with hierarchical relationship together constitute HSB multi-scale feature extractor.

HSB contains two main operations split and concatenate: split is used for feature grouping and makes the two groups after grouping have the same number of channels. When the number of features to be grouped is odd, the number of channels after split is different. Part of the two groups after separation is directly used as output, which is equivalent to identity mapping, The other part can be used as the input of the next layer convolution for deeper and more detailed feature extraction; Concatenate integrates features with the same size but different contents, so that features with different convolution degree can interact with each other. Concatenate adopts a simple superposition operation, which can better ensure the representation ability of the original features.

Following figure 1 to represent the structural schematic of HSB, HSB processes features in a convolutional layer of 3×3, Input features were split into s group x_i after convolution by 1×1, each set of features has the same number of channels, first x passes through 3×3 of convolution layer $F_i()$, the obtained results for y_i ,splitting y_i into $y_{i,1}$ directly joins the layer output x_{i+1}, as shown by the blue colored feature of the uppermost layer, $y_{i,2}$ turned into yellow feature split into two groups after the convolution operation, one group joined the output x_{i+1} of the present layer, the other group was sent into the convolution layer after the stitching operation with $y_{i,3}$, $y_{i,3}$ was equally divided into two groups of peach red feature, one group joined the output x_{i+1} of the present layer, the other group fed into the convolution layer after the stitching operation with $y_{i,4}$ to get the green feature, $y_{i,4}$ was put into the convolution layer after the same operation with $y_{i,3}$, The last set of $y_{i,5}$ features after convolution is the last part of this layer's output. After such continuous processing of features is equivalent to more scale and deeper convolution, the small receptive field in the ultimately output feature is able to focus on the detail part, enhancing the network's ability to process fine features.

ISSN: 2664-9640

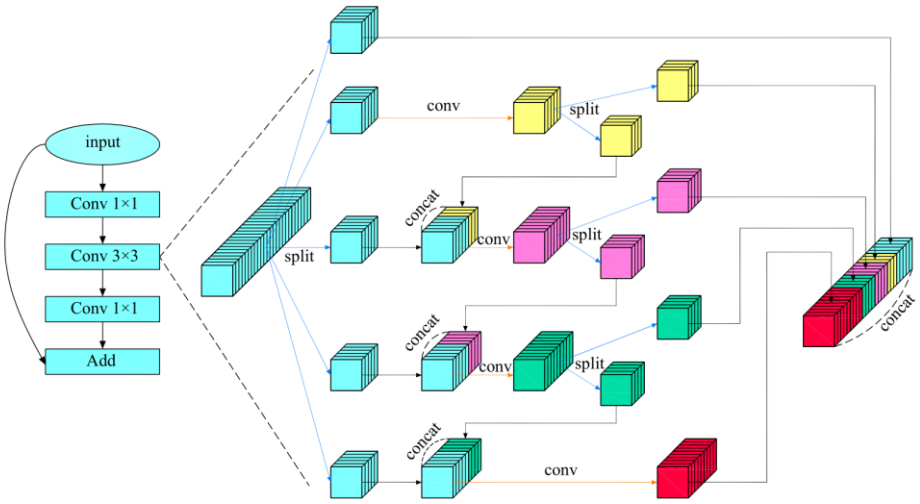


Fig. 1 HSB module

In the figure is a scenario where s is set to 5 and indeed larger group sizes enable extraction performance on more scales, with a larger number of channels implying a richer set of features and also the need to use more parameters, thus requiring a trade-off between the number of parameters and feature extraction power.

$$y_i = \begin{cases} x_i & i = 1 \\ f_i\left(x_i \oplus y_{i-1,2}\right) & 1 < i \le s \end{cases} \tag{1}$$

HSB does not make the network more parameter rich, and it has even fewer parameters than standard k×K convolutions. The parameter complexity of the standard is formulated as follows:

$$P_{normal} = k \times k \times s \times w \times s \times w = k^2 \times s^2 \times w^2 \tag{2}$$

The complexity of HSB is shown in the following equation:

$$P_{HSB} = \begin{cases} 0, & i = 1 \\ k^2 \times w^2 \times \left(\dfrac{2^{s-1}-1}{2^{s-1}} + 1\right) & 1 < i \le s \end{cases} \tag{3}$$

As can be seen from the comparison of equation (2) with (4), HSB complexity analysis is actually smaller than ordinary convolution.

$$k^2 \times w^2 \times \left(\frac{2^{s-1}-1}{2^{s-1}} + 1\right) \le k^2 \times w^2 \times \left(\frac{2^{s-1}-1}{2^{s-1}} + s - 1\right) < k^2 \times w^2 \times (s - 1 + s - 1) = k^2 \times w^2 \times (2s - 2) < k^2 \times w^2 \times s^2 \tag{4}$$

2.2. Multiscale decoder

The decoder is the encoder's deconvolution procedure with the goal of reducing the image to the original image size and the decoder samples the image up with 3 × 3 deconvolution, restoring each layer image to the same size as the decoder, with an output feature dimension of 256, 128, 64, 32, 16 per convolution.

The bilinear sampling has gradient localization and may not converge to the global minimum during the training process for the final difference estimation, so the model extracts the difference in the decoder's last four layers and computes the loss function separately by integrating it into the final solution, where the layers compute the loss function separately according to different image sizes. Because low resolution image compression is severe, it is difficult to retain image important details. Prone to the problem of discontinuity of parallax in weak repetitions of tissue structures due to the fact that photometric errors at these locations

ISSN: 2664-9640

are ambiguously inaccurate, inspired by binocular stereovision, this paper improves the loss function, reconstructs the parallax separately in the last four layers with different image sizes of the decoder, and computes the loss function at different scales separately. This paper does not directly calculate photometric error on a small-size image, but first samples the small-size low resolution parallax map to high resolution before calculating projected photometric error. The parallax in each scale was taken as the standard with high-resolution images such that the parallax in each scale was adjusted in the same direction, thus enabling effective restraint of accuracy in the parallax.

3. loss function

the fundamental loss function for unsupervised monocular depth estimation is the photometric reprojection loss. Motivated by [13], the model-driven smoothness loss is appended to better explore the solution space of the disparity over the training stage instead of the standard smoothness loss.

3.1. Photometric loss

We formulate a self-supervised signal from the image formation process via the photometric loss. Structured similarity (SSIM)[14] is a commonly-used metric for evaluating the quality of image predictions, which is adopted to measure the similarity between two image patches x and y, and can be written as:

$$\mathrm{SSIM(x,y)}=\frac{(2\mu_x\mu_y+c_1)(2\sigma_{xy}+c_2)}{(\mu_x^2+\mu_y^2+c_1)+(\sigma_x^2+\sigma_y^2+c_2)} \tag{5}$$

where μ_x, σ_x are the local means and variances over the pixel neighborhood with $c_1=0.01^2$ and $c_2=0.03^2$. Similar to [15,16] , the photometric loss is composed of L1-norm of the discrepancy between the synthesized and real images and SSIM, which is formulated as:

$$\rho(I_t,I_s^w)=\alpha\frac{1-SSIM(I_t,I_s^w)}{2}+(1-\alpha)\left\|I_t-I_s^w\right\|_1 \tag{6}$$

where the constant α is commonly set to 0.85.

Fortunately, the occluded and is-occluded regions result in the pixels from the target image not appearing in both the previous and next frames. Thus, the pixel-level minimum trick[17] is utilized to handle this problem instead of averaging the discrepancy errors from the source images. Our final photometric loss can be formulated as:

$$L_{ph}=\sum_l \min \rho_l(I_t,I_s^w) \tag{7}$$

3.2. Smoothness loss

To encourage the disparity outputs to be locally smooth meanwhile preserving sharp edge in the discontinuous regions, the edge-aware smoothness loss is usually adopted in self-supervised depth estimation:

$$L_{sm}=\left|\partial_x d_t^*\right|e^{-\left\|\partial_x I_t\right\|_1}+\left|\partial_y d_t^*\right|e^{-\left\|\partial_y I_t\right\|_1} \tag{8}$$

where $d_t^*=d_t/\overline{d_t}$ is the normalized disparity to remove the shrinking of predicted depth maps[18]. Furthermore, a spatial (pixel-level) and temporal(training-time) model-driven weight [14] will be adopted to better search the predicted depth space, and it can be written as:

ISSN: 2664-9640

$$\beta_i=\exp(-\frac{c\left\|I_t(i)-I_s^w(i)\right\|_1}{\frac{1}{N}\sum_{i=1}^{N}\left\|I_t(i)-I_s^w(i)\right\|_1}) \tag{9}$$

where N is the pixel number of the target image and c is empirically set to 10 for adjusting the range of β_i for the pixel i. Thus, our model-driven smoothness loss can be formulated as:

$$L_{md}=\frac{1}{N}\sum_{i=1}^{N}\beta_i L_{sm}^i \tag{10}$$

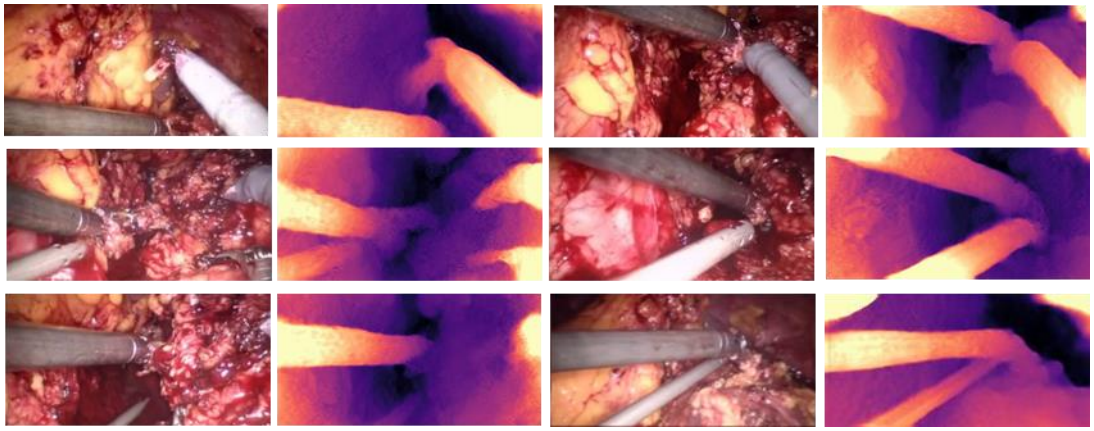
The total loss function is composed of two terms:

$$L=L_{ph}+\lambda L_{md} \tag{11}$$

4. Summary

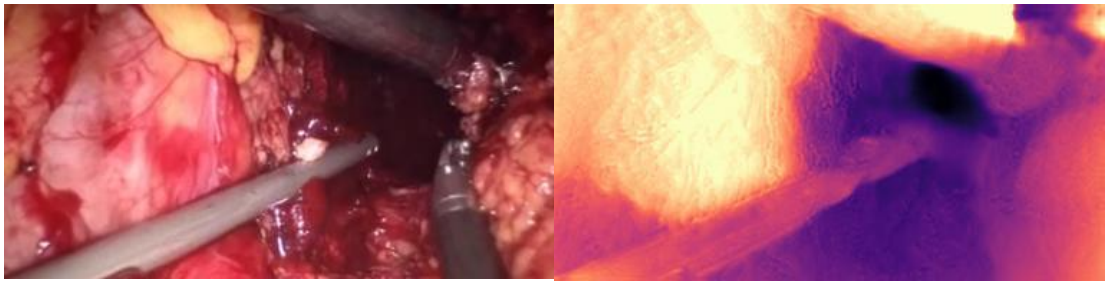
The training time in this paper is 7-8 hours. Hs-Resnet50 is used for training. The final loss of resnet50 training is 0.06, while the final loss of hs-resnet50 is about 0.05. There is no fitting between the two schemes. The loss of hs-resnet50 is low and the model training effect is good.

In Figure 2(a) and (c) represent the endoscope test image, (b) and (d) represent the RGB parallax map obtained from the HS-Resnet model in this paper. From the test image, it can be seen that the parallax map generated by this model is complete and continuous without cavity phenomenon.



(a) (b) (c) (d)

Figure. 2 disparity estimation results



(a) (b)

Figure. 3 details of parallax estimation

As shown in Figure 3, the binocular endoscope depth estimation algorithm based on HS-Resnet model can not only effectively obtain the parallax map, but also retain the image details. (a) the blood vessels in the abdominal wall in the map are well preserved in the parallax map (b). The original tissue structure and texture of the image can be observed through the parallax map,

ISSN: 2664-9640

and the blood vessel information is very important in medical images, Highlighting blood vessels and more details can also prevent doctors from accidentally injuring patients.

Since there are few literatures on disparity estimation of endoscope image and the evaluation index is not unique, this paper refers to several literatures with similar research objectives as the standard. The basic scheme of literature [19] adopts DeConvNet as the basis of model network, adopts self-monitoring scheme, takes the disparity map obtained by training endoscope image and the original image as the comparison standard, and takes the structural similarity SSIM as the index, In this paper, the same method is used to compare with it, and the results are shown in table1. Document [20] triangulates the matching points of binocular image, which makes the surrounding points easier to match and better realizes parallax estimation. Document [21] proposes a new objective optimization algorithm to solve the occlusion problem, which retains the connectivity of image segments, The boundary length is used for regularization, and finally the image segmentation and parallax estimation of natural scene images are realized. The Siamese scheme in document [19] is the structure of binocular automatic codec, and the codec structure of monocular input is basic. The initial parallax map is obtained from the structure of coder decoder, and then the virtual view is obtained from STN network, The loss is obtained by comparing the difference between the real view and the virtual view, and the appropriate parameters are obtained for each layer of the model by refining the loss function. The Siamese result obtained by binocular image is better than the basic result obtained by monocular image. The SSIM effect obtained in this paper is 0.726 ± 0.085, which is better than that obtained by Siamese.

Table 1 SSIM comparison

Model	ELAS[20]	SPS[21]	Basic[19]	Siamese[19]	Ours
Mean SSIM	0.473	0.547	0.555	0.604	0.726
Std SSIM	0.079	0.092	0.106	0.106	0.085

Similar to this paper, document [22] also adopts the structure of encoder and decoder, inputs monocular endoscope image for unsupervised training, and obtains parallax image during test. Since the data set itself does not contain real parallax value, the parallax value obtained by the current best method is used as the standard, and the parallax image obtained by SGBM in the traditional method is used as the real value in this paper, The predicted parallax value is compared with the real value, and SSIM and PSNR are taken as the standards. The results are shown in table 2. The experimental results in this paper are 0.8826 ± 0.0678 in SSIM and 17.2594 ± 1.6254 in PSNR, which are better than other methods, confirming the effectiveness of the scheme in this paper.

Table 2 comparison between PSNR and SSIM

Model	Basic	Autoencoder[22]	Ours
Mean SSIM	0.5414±0.0709	0.8349±0.0523	0.8826±0.0678
Mean PSNR	7.7650±1.3686	14.4957±1.9676	17.2594±1.6254

Acknowledgements

National Natural Science Foundation of China key project+69735101; Chongqing Municipal Education Commission Science and Technology Research Project+KJQN201800614.

References

[1] Xu Zhong, Liu Hongying, Pi Xitian. Research on Medical Ultrafine Endoscope System [J]. Chinese Journal of Biomedical Engineering, 2014, 33(01):107-111.

[2] Fan Shanhui, Liu Shichen, Cao E. Automatic identification of small intestinal polyps in wireless capsule endoscopy images [J]. Chinese Journal of Biomedical Engineering, 2019, 38(5):522-532.He Kaiming, Sun Jian, and Tang Xiaoou et al. Single image haze removal using dark channel prior[J]. IEEE Transactions on Pattern Analysis and machine Intelligence, 2011, 33(12): 2341-2353.

[3] Ronneberger O, Fischer P, Brox T, et al. U-Net: Convolutional Networks for Biomedical Image Segmentation[C] //Medical Image Computing and Computer Assisted Intervention, 2015: 234-241.

[4] Skinner K A, Zhang J, Olson E A, et al. UWStereoNet: Unsupervised Learning for Depth Estimation and Color Correction of Underwater Stereo Imagery[C] //International Conference on Robotics and Automation. Montreal, Canada: IEEE Press, 2019: 7947-7954.

[5] Zhou T, Brown M, Snavely N, et al. Unsupervised learning of depth and ego-motion from video[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1851-1858.

[6] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.

[7] Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3828-3838.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] Shu C, Yu K, Duan Z, et al. Feature-metric loss for self-supervised learning of depth and egomotion[C]//European Conference on Computer Vision. Springer, Cham, 2020: 572-588.

[10] Guizilini V, Ambrus R, Pillai S, et al. 3d packing for self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 2485-2494.

[11] Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020, 43(10): 3349-3364.

[12] Wong A, Soatto S. Bilateral cyclic constraint and adaptive regularization for unsupervised monocular depth prediction[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 5644-5653.

[13] Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity[J]. IEEE transactions on image processing, 2004, 13(4): 600-612.

[14] Godard C, Mac Aodha O, Brostow G J. Unsupervised monocular depth estimation with left-right consistency[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 270-279.

[15] Pilzer A, Lathuiliere S, Sebe N, et al. Refine and distill: Exploiting cycle-inconsistency and knowledge distillation for unsupervised monocular depth estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 9768-9777.

[16] Godard C, Mac Aodha O, Firman M, et al. Digging into self-supervised monocular depth estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 3828-3838.

[17] Wang C, Buenaposada J M, Zhu R, et al. Learning depth from monocular videos using direct methods[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 2022-2030.

[18] Ye M, Johns E, Handa A, et al. Self-supervised siamese learning on stereo image pairs for depth estimation in robotic surgery[J]. arXiv preprint arXiv:1705.08260, 2017.https:// arxiv.org /abs/1705.08260

[19] Geiger A, Roser M, Urtasun R. Efficient large-scale stereo matching[C]//Asian conference on computer vision. Springer, Berlin, Heidelberg, 2010: 25-38.

[20] Yamaguchi K, McAllester D, Urtasun R. Efficient joint segmentation, occlusion labeling, stereo and flow estimation[C]//European Conference on Computer Vision. Springer, Cham, 2014: 756-771.

ISSN: 2664-9640

[21] Xu K, Chen Z, Jia F. Unsupervised binocular depth prediction network for laparoscopic surgery[J]. Computer Assisted Surgery, 2019, 24(sup1): 30-35.