Flow-Unet for High Dimensional Image Semantic Segmentation

Qiu Xiao-meng[1,2], Wang Lin[3], Gu Wen-jun[1,2], Tian Hao-lai[4], Hu Yu[4], Song Wei[1]

[1] Henan Academy of Big Data, Zhengzhou University
[2] School of Computer and Artificial Intelligence, Zhengzhou University
[3] Beijing Weimai Medical Equipment Co.,Ltd
[4] The Institute of High Energy Physics of the Chinese Academy of Sciences

E-mail: huyu@ihep.ac.cn

Abstract: The achievement of image semantic segmentation shows the potential of the Convolutional Neural Network (CNN) for medical image analysis. However, the application of the existing CNN model to the video neglect the correlation between frames of the video. A video semantic segmentation framework based on U-Net is proposed in this article where the feature map of the previous-frame is propagated to the next frame via an optical flow field, which is called Flow-Unet. The framework includes three parts: 1) a segmentation sub module using U-Net to segment the current frame; 2) an optical flow feature extraction module to perform feature extraction on the motion information of the current frame and the previous frame; 3) a correction module, which assigns weights to the segmentation results and optical flow features to achieve the correction effect. The effectiveness of our proposed method is presented on two public datasets (Drosophila melanogaster electron micrographs, Chaos), and private Digital Subtraction Angiography (DSA) video datasets.

1. Introduction

Medical image segmentation[1,2] not only extracts regions of interest and measures human organs but also provides raw data for 3D reconstruction of medical images. However, due to factors such as imaging equipment and patients' body movement, medical images inevitably show artifacts and noise. These factors have caused certain problems and challenges for image segmentation and medical diagnosis, so it is important to study medical image segmentation methods to find better segmentation effects.

Traditional medical image segmentation[3] methods can lead to voids in the segmented region and are sensitive to noise. In order to solve these problems, researchers apply deep learning methods to image segmentation tasks, using the learning function of the relevant network to weaken the effect of noise on segmentation, thus improving its performance. The gold standard for medical image segmentation is U-Net[4], which is proposed by Ronneberger and others. Its Encoder-Decoder and skipping connection structures sufficiently fuse the information between different scales to obtain more robust segmentation results. Its morphological variations, such as TransUnet[5] and Unetr[6], have achieved good segmentation results. However, the above architectures target single-frame images, while for video streams, temporal information should be passed to improve segmentation accuracy. One of the most typical methods is optical flow[7,8], used by architectures such as Netwarp[9] and Low-latency[10].

However, the blood vessels in medical images are relatively thin, and the segmented part often has low contrast with the surrounding tissues, causing problems such as poor edge extraction and broken blood vessels in the process of segmenting medical images. In single-frame semantic segmentation, the U-Net network solves the edge extraction problem well with its unique network structure, but does not make full use of the a priori knowledge in the temporal information, so it cannot further improve the segmentation accuracy. Meanwhile, the existing video semantic segmentation [11] models utilize the temporal information, but lack in edge extraction. In this article, we proposes Flow-Unet that takes into account both medical image features and temporal information by combining optical flow, U-Net and Inception[12] structure. The model first uses optical flow to obtain motion information between two adjacent frames, then uses U-Net and Inception structure to extract features from the current frame and optical flow information respectively, and finally uses the correction module to assign weights to the features of the current frame and optical flow information to realize the correction effect of optical flow on the current frame, so as to improve the segmentation effect of the image. The experiments on the relevant datasets show that Flow-Unet achieves better prediction results than classical segmentation model.

2. Model

2.1. Dataset

In order to comprehensively study the model performance, three representative data sets of different types are selected for experiments in this article. For each of them, the image size is set to 512×512 in the pre-processing stage, and each pair of adjacent frames is divided into a group by using two-by-two grouping.

1) Coronary angiogram: It is the real dataset of a medical company, and each image has corresponding labels for the background and blood vessels, black area is background and white area is blood vessel. There are 1200 images in the training set and 38 images in the validation set.

2) Drosophila electron microscopy images: It is a public dataset provided by the ISBI challenge, and it contains 30 consecutive sets of images. Each image is accompanied by a corresponding labeled segmentation map, where white area is cell and black area is membrane.

3) CHAOS: It is a publicly available dataset consisting of abdominal contrast CT and abdominal MR contrast images, and only their CT images are used in this experiment, and the data format is DICOM. 2050 images are in the training set and 266 data in the validation set.

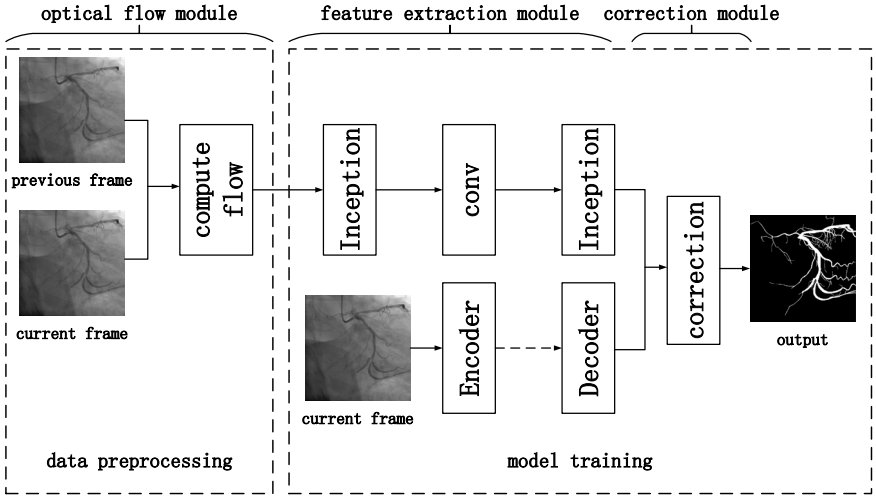

Figure 1 model architecture

2.2. Architecture

The network model proposed in this article is shown in Figure 1, including two parts: pre-processing and model training.

The pre-processing part corresponds to the optical flow module, which is used to obtain the motion information between two adjacent frames. Optical flow is the method used to describe the motion of objects in a scene that produces dynamic changes between two consecutive frames, which is essentially a two-dimensional vector field. In this article, we use TV-L1 to solve the optical flow information. Assuming that the two adjacent frames are $I0$ and $I1$, $X=(x,y)$ is the pixel points on $I0$, the energy function of TV-L1[13,14] is shown in the following equation.

$$E=\int_{\Omega}\left\{\lambda\left|u\nabla I_1+I_1(X+U_0)-U_0\nabla I_1-I_0\right|+\left|\nabla U\right|\right\}dx$$

where: $U=(u,v)$ is the two-dimensional optical flow field, ∇u and ∇v are the two-dimensional gradient, and λ is the weight constant. The former term is the data constraint term, which represents the difference in gray value between two adjacent frames at the same pixel point. While the latter term is the motion regularization constraint, which assumes that the motion is continuous.

The model training part includes the feature extraction module and the correction module. The feature extraction module uses two ways to extract features from the current frame and optical flow information to obtain the preliminary segmented image. One is to segment the current frame with U-Net, which extracts the edges well. The second is the segmentation of optical flow information by Inception structure, in which the 1×1 convolution kernel not only realizes dimensionality reduction, but also effectively reduces the number of parameters.

In addition, the segmentation information of the current frame and the optical flow will play different roles in the final segmentation result, so they are given different weights in this article . Firstly, the above two segmentation results are input to the linear layer to obtain a weight matrix, then this matrix is normalized, and finally the initial two segmentation results are multiplied by their own weight matrices to obtain the final segmentation results.

2.3. Loss function

The loss function used in this work is defined as the following equation.

$$Loss_{total}=0.5\times Loss_{dice}+0.5\times Loss_{bce}$$

where: $\text{Loss}_{dice}=1-\frac{2\left|X\cap Y\right|}{\left|X\right|+\left|Y\right|}$, $\text{Loss}_{bce}=-\sum_{i=1}^{n}y_i\log(x_i)$, X and x_i denote the predicted value, Y and y_i denote the true value, $\left|X\cap Y\right|$ denotes the dot product of the corresponding elements of the two sets, and n indicates the number of categories.

In this article, we use both loss functions, Dice and BCE. The Dice loss function focuses on similarity, which can optimize the segmentation details and improve the segmentation accuracy, and the BCE loss function makes the pixels maintain smooth gradients.

3. Experiments and analysis of results

3.1. Coronary angiogram segmentation results

The task of this dataset is to segment out the vascular information, and in order to verify the reliability of the model in this article, six classical models and Flow-Unet are selected for comparison tests in the same experimental setting. Meanwhile, three samples are randomly selected and their prediction results on U-Net, U-Net++[15], AttentionUnet[16] and RefineNet[17] and Flow-Unet are presented, and the prediction results are shown in Figure 2.

The red labeled boxes in the figure indicate the cases where the vessels are broken. Compared with the comparison model, the number of breaks in Flow-Unet is less and the image connectivity is better. The yellow boxes indicate noise generated by the model and the Flow-Unet model extract a richer amount of information. In summary, the best segmentation effect is achieved in this article on coronary angiogram.

3.2. Drosophila electron microscopic segmentation results

To further validate the effectiveness of Flow-Unet, experiments were conducted on the Drosophila electron microscopy dataset using the same experimental setting. Three images are also randomly selected and their predictions on U-Net, U-Net++, PspNet[18], SegNet[19] and Flow-Unet are shown in Figure 3.

The segmentation of U-Net and U-Net++ has a higher probability of breaking from the red labeled boxes in the figure, which leads to a reduction in the connectivity of the cells. The yellow labeled box is still noisy information, which shows that the comparison models, especially U-Net++, are more likely to be noisy. Meanwhile, in the green labeled box, the segmentation part does not break nor produce noise, but the original saw-shaped features of the cells cannot be identified. In contrast, the segmentation information obtained from Flow-Unet performs well in all aspects, and the segmentation is highly accurate.

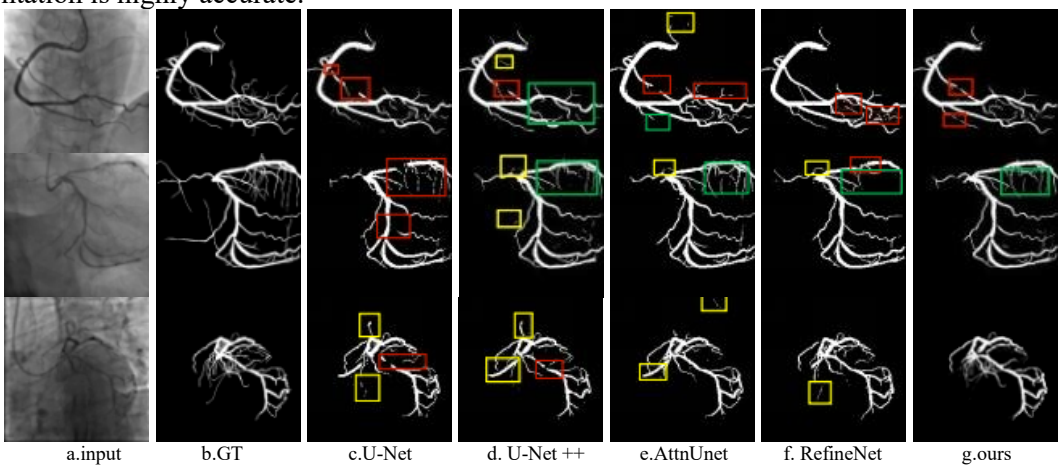

Figure 2 Segmentation results of Coronary artery

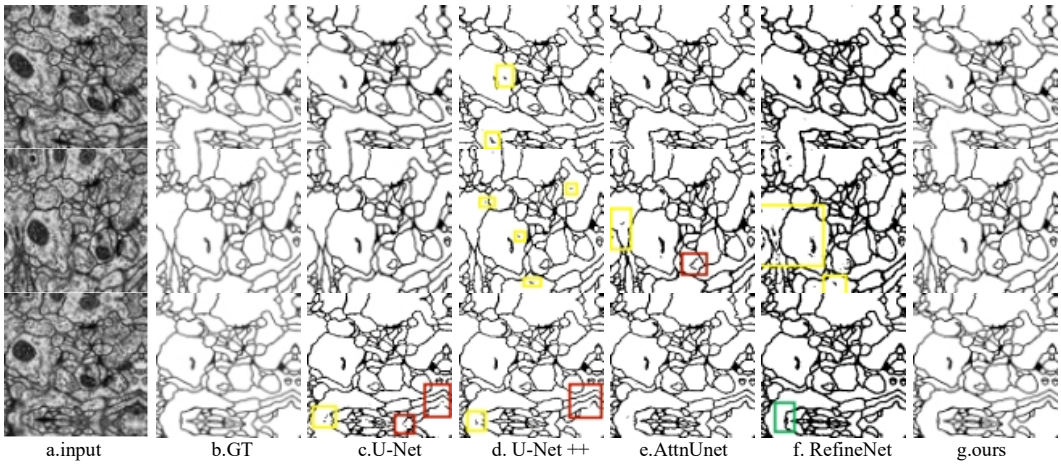

Figure 3 Segmentation results of Drosophila electron microscope

3.3. Comprehensive abdominal organ (CHAOS) segmentation results

To verify the accuracy and effectiveness of Flow-Unet and the comparison model on the CHAOS segmentation task, this article performs segmentation experiments on the CHAOS dataset with the same environment. Three images are also randomly selected and their prediction results on U-Net, U-Net++, AttentionUnet and RefineNet as well as Flow-Unet are shown in Figure 4.

In the first prediction image, it can be seen that U-Net ++ and RefineNet do not predict accurately on the protruding parts and do not capture enough information. In the second figure, AttnUnet produces small black holes in the prediction map. In the third figure, the predictions of the comparison

models both produce noise. It can be seen that Flow-Unet not only has good segmentation on the protruding parts, but also basically does not produce noisy information and has a high accuracy of segmentation.

From the segmentation results on the three data sets, it can be seen that Flow-Unet achieved best results for all of them. In order to more intuitively show the segmentation effect of the models, the evaluation indexes of each model on these three data sets were calculated in this article, and the results are shown in Table 1. As can be seen from the table, Flow-Unet improved 0.6%, 0.13% and 2.13% in the coronary angiogram, 0.42%, 0.67% and 0.83% in the Drosophila electron microscopy, and 0.88%, 0.02% and 1.68% in the CHAOS data set, respectively. This shows that Flow-Unet has good effectiveness and generalization. Meanwhile, it can be seen that U-Net and its variants tend to present relatively better results in segmentation results, so the selection of U-Net as the backbone network is a very suitable choice.

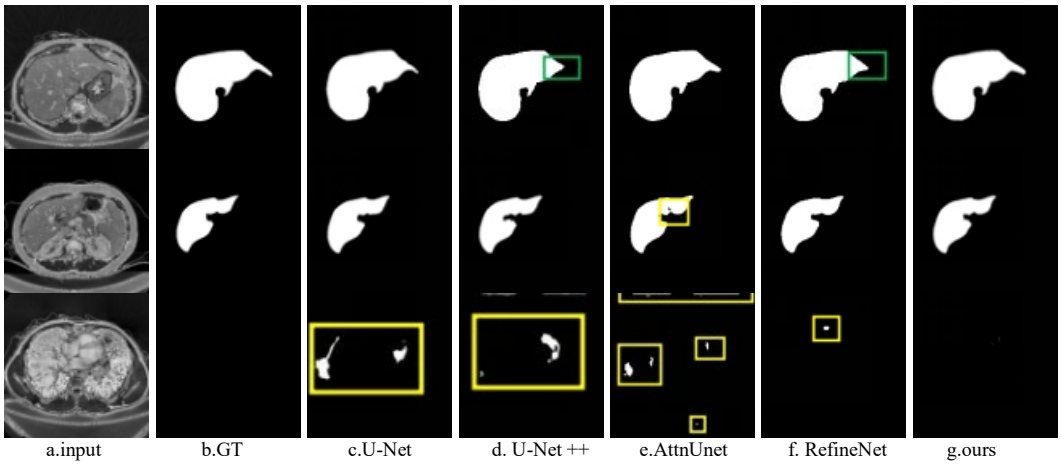

Figure 4 Segmentation results of CHAOS

Table 1 Evaluation Indicators of Comparison Test

		Dice	PA	IoU
Coronary angiogram	U-Net	0.7966	0.9915	0.8376
	U-Net++	0.8042	0.9736	0.6725
	AttentionUnet	0.8032	0.9783	0.6853
	PspNet	0.7504	0.9886	0.7725
	SegNet	0.7847	0.9711	0.6457
	RefineNet	0.7873	0.9705	0.6492
	Ours	**0.8102**	**0.9928**	**0.8589**
Electron micrograph of Drosophila	U-Net	0.9882	0.9811	0.9767
	U-Net++	0.9863	0.9780	0.9729
	AttentionUnet	0.9553	0.9306	0.9145
	PspNet	0.9830	0.9730	0.9665
	SegNet	0.9808	0.9721	0.9623
	RefineNet	0.9769	0.9661	0.9549
	Ours	**0.9924**	**0.9878**	**0.9850**
CHAOS	U-Net	0.9575	0.9953	0.9186
	U-Net++	0.9652	0.9961	0.9327
	AttentionUnet	0.9672	0.9964	0.9364
	PspNet	0.9504	0.9945	0.9055
	SegNet	0.9500	0.9944	0.9047
	RefineNet	0.9621	0.9958	0.9270
	Ours	**0.9760**	**0.9966**	**0.9532**

3.4. Parameter analysis and ablation experiments

The important parameters of a model, such as the learning rate and the optimizer, have some influence on the training results. Also, the number of Inception structures will have an impact on the results of Flow-Unet, so the three parameters are compared and experimented. In addition, considering that the two segmentation results have different degrees of influence on the final output results, a correction module is used to assign weights to the two segmentation results in Flow-Unet. In the ablation experimental section, experiments were performed without the Inception structure and correction module. The results of their experiments on the CHAOS dataset are shown in Table 2. It can be seen from the table that our models are selected with optimal parameters and each module is indispensable.

Table 2 Analysis of Important Parameters and Ablation experiment

		Dice	PA	IoU
Learning Rate	10^{-2}	0.9605	0.9956	0.9240
	10^{-3}	0.9675	0.9964	0.9371
	10^{-5}	0.9665	0.9963	0.9353
	10^{-6}	0.9343	0.9931	0.8767
Optimizer	SGD	0.7833	0.9801	0.6438
	Adagrad	0.9507	0.9946	0.9060
Number of Inception modules	Inception 1	0.9727	0.9962	0.9468
	Inception 3	0.9730	0.9962	0.9475
Ablation experiments	no-Inception	0.9579	0.9953	0.9193
	no-correction	0.9663	0.9954	0.9349
	Ours	**0.9760**	**0.9966**	**0.9532**

4. Summary

In this article, a novel semantic segmentation model for temporal images is proposed. The model uses the U-Net network as the backbone, taking full advantage of its high performance in medical image segmentation, supplemented by optical flow to transmit motion information. The effectiveness and generalization of the model proposed in this article are verified by image segmentation on relevant datasets as well as ablation experiments. The experimental results show that Flow-Unet obtains better segmentation results compared with the classical model and thus further enhances the reference value in clinical diagnosis.

In order to further optimize the segmentation effect, two issues still need to be further considered in this article: (1) how to better achieve the prediction of edges so that the segmentation results can better fit the medical target area; (2) how to reduce the noise generated by various models at the beginning of the temporal sequence map when no lesion information may appear.

References

[1] Liu X, Song L, Liu S, et al. A review of deep-learning-based medical image segmentation methods[J]. Sustainability, 2021, 13(3): 1224.

[2] Hesamian M H, Jia W, He X, et al. Deep learning techniques for medical image segmentation: achievements and challenges[J]. Journal of digital imaging, 2019, 32: 582-596.

[3] Meiburger K M, Acharya U R, Molinari F. Automated localization and segmentation techniques for B-mode ultrasound images: A review[J]. Computers in biology and medicine, 2018, 92: 210-235.

[4] 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.

[5] Chen J, Lu Y, Yu Q, et al. Transunet: Transformers make strong encoders for medical image segmentation[J]. arXiv preprint arXiv:2102.04306, 2021.
[6] Hatamizadeh A, Tang Y, Nath V, et al. Unetr: Transformers for 3d medical image segmentation[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2022: 574-584.
[7] Zhai M, Xiang X, Lv N, et al. Optical flow and scene flow estimation: A survey[J]. Pattern Recognition, 2021, 114: 107861.
[8] Luo A, Yang F, Li X, et al. Learning optical flow with kernel patch attention[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 8906-8915.
[9] Gadde R, Jampani V, Gehler P V. Semantic video cnns through representation warping[C]//Proceedings of the IEEE International Conference on Computer Vision. 2017: 4453~4462.
[10] Li Y, Shi J, Lin D. Low-latency video semantic segmentation[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 5997~6005.
[11] Wang W, Zhou T, Porikli F, et al. A survey on deep learning technique for video segmentation[J]. arXiv preprint arXiv:2107.01153, 2021.
[12] Si C, Yu W, Zhou P, et al. Inception transformer[J]. arXiv preprint arXiv:2205.12956, 2022.
[13] Mohamed N A, Zulkifley M A. Moving object detection via TV-L1 optical flow in fall-down videos[J]. Bulletin of Electrical Engineering and Informatics, 2019, 8(3): 839-846.
[14] Padmavathi K, Asha C S, Maya V K. A novel medical image fusion by combining TV-L1 decomposed textures based on adaptive weighting scheme[J]. Engineering Science and Technology, an International Journal, 2020, 23(1): 225-239.
[15] Le Duy Huynh N B. A u-net++ with pre-trained efficientnet backbone for segmentation of diseases and artifacts in endoscopy images and videos[C]//CEUR Workshop Proceedings. 2020, 2595: 13-17.
[16] Oktay O, Schlemper J, Folgoc L L, et al. Attention u-net: Learning where to look for the pancreas[J]. arXiv preprint arXiv:1804.03999, 2018.
[17] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925-1934.
[18] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-28.
[19] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.