LVC-Net: Medical Image Segmentation with Noisy Label Based on Local Visual Cues

Yucheng Shu, Xiao Wu, and Weisheng Li

Chongqing University of Posts and Telecommunications, Chongqing 400065, China
shuyc@cqupt.edu.cn wxwsx1997@gmail.com liws@cqupt.edu.cn

Abstract. CNN-based deep architecture has been successfully applied to medical image semantic segmentation task because of its effective feature learning mechanism. However, due to the lack of semantic guidance, such supervised learning model may be susceptible to annotation noise. In order to address this problem, we propose a novel medical image segmentation algorithm based on automatic label error correction. Firstly, local visual saliency regions, namely the Local Visual Cues (LVCs), are captured from low-level feature channels. Then, a deformable spatial transformation module is integrated into our LVC-Net to build visual connections between the predictions and LVCs. By combining noisy labels with image LVCs, a novel loss function is proposed based on their intrinsic spatial relationship. Our method can effectively suppress the influence of label noise by utilizing potential visual guidance during the learning process, thereby generate better semantic segmentation results. Comparative experiment on hip x-ray image segmentation task demonstrate that our algorithm achieves significant improvement over state-of-the-arts in the presences of noisy label.

1 Introduction

Semantic segmentation plays an important role in the field of medical image analysis and understanding. By extracting the visual homogeneous regions within the image, it can provide useful information to many visual tasks, such as medical image recognition, medical image registration, medical image reconstruction, etc.

With the theoretical development of machine learning techniques, most state-of-the-art methods were proposed to utilize the feature learning mechanism of deep neural networks and have shown promising medical image segmentation results [1–3]. Among them, Fully Convolutional Networks (FCN) [4] and U-Net [5] are most popular methods in this field. FCN define a skip architecture that combines semantic information from a coarse layer with appearance information from a fine layer to produce detailed segmentations. In an encoder-decoder fashion, U-Net consists of a contracting path to capture context and a symmetric expanding path that enables precise localization. Since then, more and more medical segmentation methods are proposed based on these fundamental architectures. SegNet [6] uses pooling indices computed in the max-pooling step of the corresponding encoder to perform non-linear upsampling. Md Zahangir

Alom, etc. [7] propose a Recurrent Residual Convolutional Neural Network based on U-Net model. Compared to equivalent models, this method, also known as R2U-net, shows superior performance on many medical segmentation tasks.

To sum up, the skip-connections used in such methods can directly forward low level information to the backend of network such that the segmentation accuracy can be greatly improved. Yet its effectiveness highly depends on the quality of annotations. Prior work [8] shows that neural networks have such a strong capability to converge even if the training data is inaccurate. Therefore if the annotation is greatly contaminated, the real and effective features may not be necessarily learned. In such cases, especially when the learning model is built upon a label-dependent supervised framework, there is no guarantee that the algorithm could spontaneously discover useful visual cues to fight against wrong annotations and eventually reach real optima. To address this problem, Zhiwu Lu, etc. [9] proposes a sparse learning model to directly and explicitly detect noisy labels in a superpixel setting. Anna Khoreva, etc. [10] design a weakly supervised learning technique to acquire semantic label from bounding box detection annotations. More recently, a new image boundary generation framework, namely STEAL [11], is proposed to learn more accurate semantic boundaries from noisy annotations by enforcing maximum response along the normal edge direction.

Therefore, in order to acquire meaningful segmentations under the influence of label noisy, it is essential to conduct weak-supervised learning by integrating useful prior information that sterns directly from the images. In this paper, we present a novel medical image segmentation algorithm, namely the LVC-Net, to deal with the label noise. Firstly, since the front-end of network is relatively less affected by supervised signals, we propose to capture local visual saliency features, called Local Visual Cues (LVCs), from low-level convolutional channels. Then we further integrate an deformable spatial transformation module into our encoder-decoder network, to provide extra freedom for local receptive fields so that the network could spontaneously establish visual connections between the predictions and LVCs during the learning process. Finally, a novel loss function is proposed to build effective regularizations for the noisy label, the LVCs, and the deformable modules. Comparative experiment on hip x-ray image segmentation task demonstrate that our algorithm achieves significant improvement over baselines in the presences of noisy label.

2 Method

2.1 Overview

In our framework, we employ the classic U-Net as our basic build block. The encoder-decoder net is built based on traditional convolutional layer, max-pooling layer, and up-sampling layer. Note that these basic operational layers can be easily transformed into modern recurrent or residual version as proposed in [7]. As shown in fig.1, the whole layout consists 3 parts: the U-Net module, the LVC generator, and the spatial transform module. We use a multi-loss function to

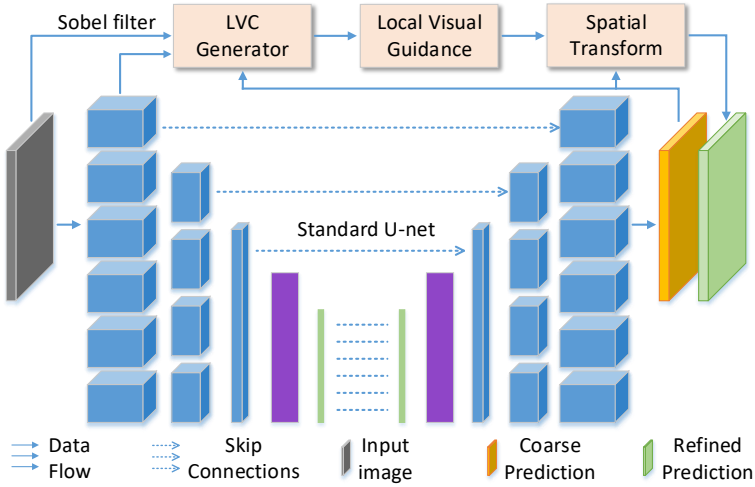


Fig. 1. The framework of LVC-Net

guide the learning behavior of our network. Detailed information is shown in the following sections.

2.2 Local Visual Cues

As discussed above, in supervised deep learning architecture, the annotation plays an essential role as it is the only training criterion that the learning algorithms have to trust. Although low-level features fed by the skip-connection may bring meaningful influences to the high-level "decision-making" branch, once the labels are badly corrupted, the neural networks still have the capability to converge or even over-fit the wrong training data[8]. In medical image segmentation tasks, the algorithms are often required to output detailed pixel-level segmentation results, but if the annotations are not trustworthy, it may bring a great challenge to most existing supervised methods. And because of the complexity and diversity of medical images, the label-generating work can be both tricky and onerous, thus the label noise of medical image is much of a practical and urgent issue.

It is a common observation that during the learning process, the front-end of network may less affected by supervised signals, thus can be used to generate detailed image features like the method proposed in [12]. Inspired by this work, we propose to capture local visual saliency features, called Local Visual Cues (LVCs), from low-level convolutional channels.

Specifically, the homogeneous pixels can be clustered into one class because of their visual and semantic similarity, thus one of the most important low-level visual cue for segmentation task is the inter-class boundaries, or image edges. To fully utilize the deep network and learn class-specific LVCs, we first use generic image filters such as Sobel operator to generate the initial coarse

ground truth. By incorporating frond-end channels, prediction results and the initial edge ground truth, the LVC generation loss is proposed to calculate the cross-entropy at every pixel as:

$$LVC_{loss} = \begin{cases} P(s) \cdot y_s \cdot \log(y_i) & \rho \leq y_i \leq 1 \\ (1 - P(s)) \cdot (1 - y_s) \cdot \log(1 - y_i) & 0 \leq y_i < \rho \end{cases} \tag{1}$$

where y_i is the output of LVC generator, $P(s)$ is the intensity ratio of Sobel signals in each initial edge ground truth mask, y_s represents the class weight, and ρ is set to a typical value of 0.5. Then, a regularization term is proposed to punish LVCs that are away from prediction boundaries. Relative spatial computations will be introduced in following section. All the computations mentioned above are preformed in a differentiable manner such that the network is able to learn and preserve useful low-level visual cues to compensate weak predictions wherever label noise appears. The LVC generator and the segmentation framework are trained alternately to have a better segmentation performance.

2.3 Spatial Transformation Based on Visual Guidance

After the extraction of LVCs, the segmentation net is able to perform classification base on both noisy annotations and the local visual cues. However, the direct computation of the pixel-wise loss is rather discrete that the LVCs may be suppressed and may cause spatial discontinuous. In order to provide extra freedom for local receptive fields so that the network could spontaneously establish visual connections between the predictions and LVCs during the learning process, we propose a spatial transform module to navigate the classification boundaries away from wrong labels.

As shown in Fig.2, the proposed spatial transform module consists of two parts. By multiply the network prediction and LVC responses with a coordinate template, the position of local visual attentions is acquired as the visual guidance. Then we perform a spatial convolution step to get 2-dimensional offsets for each pixel. To update the predictions, we create a parameterized sampling grid to cultivate spatial calculation into a differentiable way. Based on the offset between each pixel and visual attentions centers, a regularization loss is proposed to control the deformation ability of the network. Eventually, the classification probability of each pixel could be re-assigned accordingly.

Local Visual Guidance We generate a coordinate template by using 2D convolution with 1x1 kernel to get a the full sized coordinate mask:

$$R = \{(x, y) | 0 \leq x \leq W, 0 \leq y \leq H, (x, y) \in \mathbb{N} \times \mathbb{N}\} \tag{2}$$

Then, the location of local visual attention centers is calculated based on the network prediction and LVC responses. By using Eq(3), the values of prediction

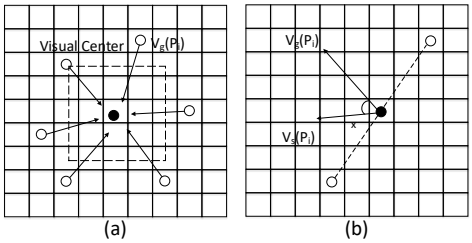


Fig. 2. The spatial transformation based on local visual guidance. (a)Local Visual Guidance acquisition. (b)Spatial Sampling Offsets. Dotted line indicates the classification boundary.

and LVCs are both taken into account, and that the network is able to estimate probabilistic centers of local visual attention.

$$\text{center}(p_i) = \frac{1}{|N(p_i)|} \sum_{j \in N(p_i)} Y\left(p_j\right) L\left(p_j\right) R\left(p_j\right) \tag{3}$$

where $Y\left(p_j\right)$ is the network prediction, $L\left(p_j\right)$ is the LVC responses, $R\left(p_j\right)$ is the coordinate mask value of p_j, $N(p_i)$ is a convolutional window, which forms a local receptive filed to investigate the relative intensity of the prediction an LVCs. If the network failed to generate correct predictions because of the label noise, the LVC in that region will take over and provide guidance for the segmentation by:

$$V_g(p_i) = \left[R\left(p_i\right) - \text{center}(p_i)\right]_{x,y} \tag{4}$$

in which V_g is the local visual guidance for each pixel to re-sample classification probability accordingly.

Spatial Sampling Offsets In order to have a better generalization ability and transformation freedom, an adaptive spatial transform model is proposed so that the segmentation sampling offsets can be inferred based on different local visual information. We adopt the spatial-convolution and sampling method proposed in [13], which applying convolution layers to learn adaptive spatial transformation. At each pixel, a offset $V_s(p_i)$ will be generated by the spatial transformation module, and then spatial sampling parameters can be learned by using the L2 regularization between sampling offsets and the local visual guidance:

$$G_{loss} = \left\| V_g(p_i) - V_s(p_i) \right\|^2 \tag{5}$$

Then the original segmentation prediction can be re-sampled based on offsets that associated with LVCs and noisy labels.

$$S(p_i) = \sum_{j \in N(p_i)} \phi\left(V_s(p_j)\right) w\left(Y(p_j)\right) \tag{6}$$

where ϕ is the standard network sampling function, and w is the local intensity weighting function.

2.4 LVC-Net losses

The final prediction of the network is made by integrating three losses: 1) LVC-generation loss LVC_{loss}, 2) LVC-guidance loss G_{loss}, and 3) LVC-segmentation loss S_{loss}. During training stage, this losses can be trained successively to shape the final segmentation LVC-Net.

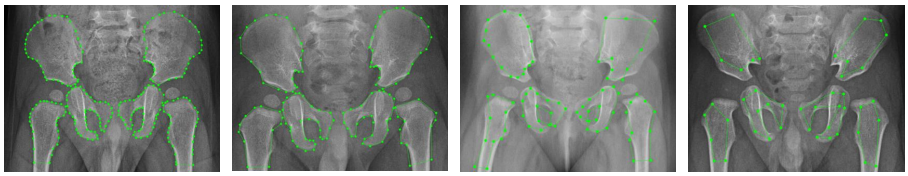
As mentioned above, we propose Local Visual Cues to provide extra visual regularizations to fight against bad annotations. Once the LVCs are obtained, the original pixel-wise segmentation loss can be re-written as:

$$S_{loss} = -\left[y_s \log \widehat{y} - y_s \log\left(1-\widehat{y}\right)\right] - \left[y_G \log \widehat{y} + \left(1-y_G\right) \log\left(1-\widehat{y}\right)\right] \quad (7)$$

where S_{loss} is the LVC-segmentation loss. The predictions will be dynamically affected by both label masks y_G and LVC masks y_s. Therefore, the final LVC-Net is able to spontaneously discover local visual information, and attempt to compensate weak predictions wherever label noise appears based on a spatial transformation mechanism.

3 Experiment

We conduct comparative experiments with FCN and U-Net on hip X-ray image data set. The convolution layer is initialized by default setup, and we use stochastic gradient descent with the learning rate of 10^{-4}. All the networks are trained for same 1200 epochs on a Nvidia Tesla v100 GPU.



(a) level 1 (b) level 2 (c) level 3 (d) level 4

Fig. 3. Label noise from level 1 to level 4

Data preparation and Evaluation metric COCO format is used to build our dataset that contains images and mask annotations. The size of all training and testing image is normalized to [128, 128], and the total number of images in each set is 1464 and 366, respectively. As show in fig.3, for each training image, we manually generated 4 groups of annotations by Labelme with the label noise form level 1 up to level 4 (higher level indicates coarser annotations). The label quality of ground truth in testing dataset is equivalent to level 1. We trained all the networks (FCN, U-Net, LVC-Net) for each label group, and evaluated the segmentation performance accordingly. The evaluation metric we use is segmentation precession, recall and the mIoU. We performed 5 rounds of evaluation.

In each round, 50 images were randomly selected, and mean evaluation metrics were calculated. Experimental results are presented in the following section.

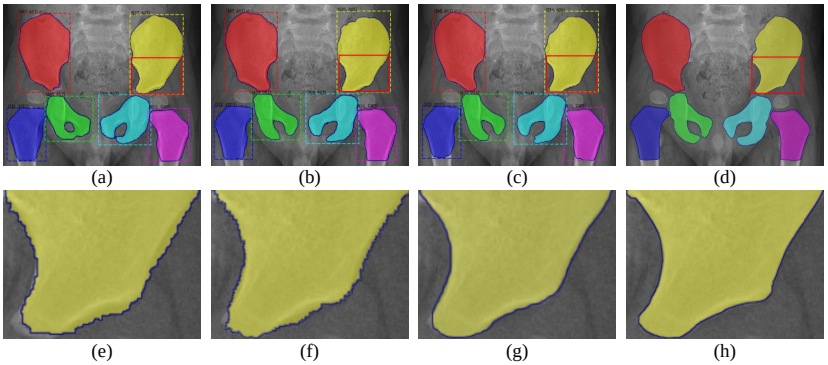


Fig. 4. Segmentation Examples. The first row is the result of different algorithms on the entire X-ray image, the second row shows detailed results within the red squares. (a)(e)FCN,(b)(f)U-Net,(c)(g)LVC-Net,(d)(h)Ground Truth

Results Some segmentation results are presented in Fig.4 in which all the nets were trained in the level 3 noisy deta set for same epochs. As clearly shown in this picture, compared to other networks, our LVC-Net is capable of discovering more accurate edges of left ilium bone under the guidance of local visual cues generated directly from the image. In contrast, the segmentation precision of FCN and U-Net are greatly affected by bad annotations. More statistical results are presented in Table.1. The results indicate that when the label noise is gradually increased, the performance of FCN and U-Net drops at a faster rate. However, our proposed method remains relatively robust at higher level of label noise. We also did module experiments on proposed methods, results demonstrate that our proposed LVC-generation module and LVC-based spatial transform module is able to generate meaningful visual information and further improver the medical image segmentation performance. And because of the space limitation, all experiment results will be presented in our extended journal version.

4 Conclusion

In this paper, we introduced a novel medical image segmentation network based on Local Visual Cues. Firstly, the LVC generator is proposed to capture local visual saliency features from low-level convolutional channels. Then, a deformable spatial transformation module is integrated into our LVC-Net to provide extra freedom for the network to build visual connections between the predictions and

Table 1. Statistical results under different levels of label noise

Noise Level	Methods	Precision	Recall	mIoU
level1	FCN	0.894	0.924	0.831
level1	U-Net	0.912	**0.955**	0.874
level1	LVC-Net	**0.951**	0.929	**0.886**
level2	FCN	0.811	0.878	0.770
level2	U-Net	0.835	0.883	0.804
level2	LVC-Net	**0.915**	**0.906**	**0.836**
level3	FCN	0.784	0.816	0.731
level3	U-Net	0.790	0.828	0.772
level3	LVC-Net	**0.902**	**0.859**	**0.810**
level4	FCN	0.652	0.704	0.620
level4	U-Net	0.664	0.728	0.690
level4	LVC-Net	**0.815**	**0.796**	**0.738**

LVCs during the learning process. The multi-loss function is proposed to build effective regularizations. Experiments on hip medical image data sets indicated that the proposed method can effectively suppress the influence of label noise, thus outperformed the state-of-the-art on segmentation tasks.

5 Acknowledgments

This research was funded in part by National Key R&D Program of China 2016YFC1000307-3, National Natural Science Foundation of China 61801068, Natural Science Foundation of Chongqing cstc2016jcyjA0407, Scientific and Technological Research Program of Chongqing Education Commission KJ1600419. The authors would like to thank Prof. Guoxin Nan for providing the data.

References

1. Yu, Changqian, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Learning a discriminative feature network for semantic segmentation. In IEEE Conference on Computer Vision and Pattern Recognition, 1857–1866 (2018)
2. Kaul, Chaitanya, Suresh Manandhar, and Nick Pears. FocusNet: An attention-based Fully Convolutional Network for Medical Image Segmentation. arXiv preprint arXiv:1902.03091 (2019)
3. Milletari, Fausto, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In 2016 Fourth International Conference on 3D Vision (3DV), 565–571 (2016)
4. Long, Jonathan, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In IEEE conference on computer vision and pattern recognition, 3431–3440 (2015)
5. Ronneberger, Olaf, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, 234–241 (2015)

6. Badrinarayanan, Vijay, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence 39(12), 2481–2495 (2017)
7. Alom, Md Zahangir, Mahmudul Hasan, Chris Yakopcic, Tarek M. Taha, and Vijayan K. Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation, arXiv preprint arXiv:1802.06955 (2018)
8. Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization, arXiv preprint arXiv:1611.03530 (2016)
9. Lu, Z., Fu, Z., Xiang, T., Han, P., Wang, L., Gao, X. Learning from weak and noisy labels for semantic segmentation. IEEE transactions on pattern analysis and machine intelligence 39(3), 486–500 (2016)
10. Khoreva, A., Benenson, R., Hosang, J., Hein, M., Schiele, B. Simple does it: Weakly supervised instance and semantic segmentation. In IEEE conference on computer vision and pattern recognition, 876–885 (2017)
11. Acuna, David, Amlan Kar, and Sanja Fidler. Devil is in the edges: Learning semantic boundaries from noisy annotations. In IEEE Conference on Computer Vision and Pattern Recognition, 11075–11083 (2019)
12. Y. Liu, M. Cheng, X. Hu, K. Wang and X. Bai, Richer Convolutional Features for Edge Detection. In IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, 5872–5881 (2017)
13. J. Dai et al., Deformable Convolutional Networks. In IEEE International Conference on Computer Vision, 764–773 (2017)