Walking on Two Legs: Learning Image Segmentation with Noisy Labels

Guohua Cheng[*]
Fudan University
Shanghai, China
17110850005@fudan.edu.cn

Hongli Ji
Jianpei Technology Co.,Ltd
Hangzhou, China
ji.hongli@jianpeicn.com

Yan Tian[†]
Zhejiang Gongshang University
Hangzhou, China
tianyan@zjgsu.edu.cn

Abstract

Image segmentation automatically segments a target object in an image and has recently achieved prominent progress due to the development of deep convolutional neural networks (DCNNs). However, the quality of manual labels plays an essential role in the segmentation accuracy, while in practice it could vary a lot and in turn could substantially mislead the training process and limit the effectiveness. In this paper, we propose a novel label refinement and sample reweighting method, and a novel generative adversarial network (GAN) is introduced to fuse these two models into an integrated framework. We evaluate our approach on the publicly available datasets, and the results show our approach to be competitive when compared with other state-of-the-art approaches dealing with the noisy labels in image segmentation.

1 INTRODUCTION

Image segmentation is a fundamental and challenging problem that aims to separate the objects and the background pixels in a given image. It has been an active area of research in computer vision over the past decade, with a wide range of applications. Examples of applications are image editing (Wang, 2018), media diagnosis (Tian, 2019a), and autonomous driving (Chen, 2018).

The development in the deep learning methods greatly promotes the effectiveness in the image segmentation. However, they require a large amount of manually labeled data. Trusting labels with noise as ground truth by encoding them as hard labels can lead to overconfident mistakes and propagated errors. Examples of noisy labels are illustrated in Fig. 1. The top, middle, and bottom rows show the input images, the noisy labels, and the correct labels, respectively. Note that the leftmost three samples illustrate that the cheek is easy to be annotated as the gum, and the rightmost sample shows the white tongue guard has a high similarity with the tooth.

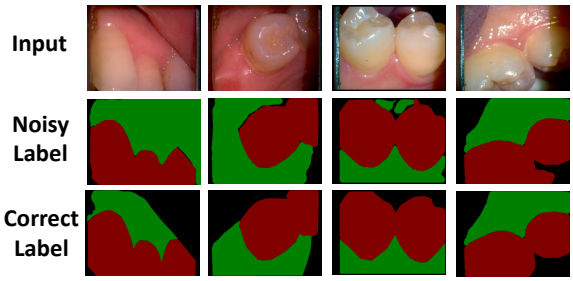

Figure 1: **Illustrations of the annotation challenge in image segmentation. The top, middle, and bottom rows show the input images, the noisy labels, and the correct labels, respectively.**

Most approaches (Kaneko, 2019) model the probability of the noisy labels and learn the noise transition matrix. These approaches assume that there is a single transition probability between the noisy label and the ground-truth label, and this probability is independent of individual samples. But in real cases, the appearance of each sample has much influence on whether it can be misclassified. Moreover, the noise transition matrix cannot be modeled well in image segmentation because of the high dimension of the annotated mask. Recently, some experts begin to directly learn the mapping between the noisy labels and the clean labels using the deep convolutional neural networks (DCNNs) (Li, 2019). However, the image features only work as a visual clue with a reduced dimension, providing neither the semantic nor the position information to benefit the label correction.

[*]This is the corresponding author.

[†]Yan Tian also works with Shining3D Tech Co., Ltd., Hangzhou, China.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

We are motivated by the fact that although the revised annotations have been improved, they still contain somewhat noises due to the ambiguous factors hidden in the data. If they are equally treated as the real annotations, the segmentator to be learned will be confused and decrease the convergence speed. We focus on how to revise the annotation and measure the revision quality in the training stage. We need to design a mechanism to combine the label correction, the sample reweighting, and the image segmentation into an integrated framework.

In this paper, we propose a generative adversarial network (GAN)-based approach to generate the revised annotation from the noisy label and use the discriminator output to obtain the weighted loss function, decreasing the influence of the untrusted samples. The novelty of this approach is the following:

- We propose a GAN-based framework to integrate the label correction and sample reweighting to improve overfitting in the training phrase caused by the noisy labels.
- We propose a recurrent neural network that explores the local context to learn the long-term dependency, and progressively propagates the visual information to refine the annotated mask.
- We introduce the discriminator into the noisy label reweighting, using the confidence obtained from the discriminator to adjust the weight in the loss function.

Experiments on the PASCAL VOC 2012 and the Shining3D dental datasets show that the proposed approach is competitive when compared with state-of-the-art approaches handling the noisy labels in segmentation.

The rest of this paper is organized as follows. Section 2 reviews studies on image segmentation and noisy label correction. Section 3 presents our GAN-based approach for jointly label correction and sample reweighting. Section 4 presents experimental results. Concluding remarks appear in Section 5.

2 RELATED WORK

Here we briefly review the literature on image segmentation and noisy label correction. We present advantages and drawbacks of each kind of approach.

2.1 IMAGE SEGMENTATION

Image segmentation learns the complex nonlinear regression from the visual space to the segmented mask. Its effectiveness improves a lot due to the development of the DCNNs. Most approaches use the encoder-decoder structure (Chen, 2018) as the encoder extracts the global feature and the decoder generates both the semantic and detail results according to the guidance from the global feature. Multiscale analysis is introduced to jointly learn discriminant features from different fields, for example, the image pyramids (Orsic, 2019), the deconvolution network (Noh, 2015), the hyperfeature (Tian, 2018), the aggregated residual block (Tian, 2019b), the dilated convolution(Chen, 2018), the multiscale attention (tian, 2019a), and the index network(Lu, 2019). The context exploitation is also important in image segmentation due to providing additional information for ambiguous situations. The context knowledge includes but not limited to the global feature (Takahama, 2019), the boundary constraint (Tian, 2017), the local similarity (Jiang, 2018), the long-term dependency (Song, 2019), and the relation between objects (Wang, 2017) or frames (Tian, 2019c). However, the supervised learning is sensitive to the annotation quality. If the labels contain noises, the decision boundary to be learned is affected and deviates the optimum solution. Moreover, the deep learning method requires amount of the training samples due to the millions of weight parameters to be learned, and the lack of data leads to the overfitting problem.

2.2 NOISY LABEL CORRECTION

Many approaches have been proposed to handle the noisy labels in recent years. Epistimic uncertainty represents the model uncertainty and labels uncertainty coming from the noisy labels that are jointly learned to explore the relation between each other (Tomczack, 2019). The sample selection-based approaches (Tu, 2020) distinguish the noisy labels according to the relevance between the samples. For example, the local and global consistency learning (LLGC) between the superpixels is proposed to optimize the segmented mask (Li, 2019). However, since the data for training are selected on the fly rather than selected in the beginning, it is hard to characterize these sample-selection biases, and then it is also difficult to give any theoretical guarantee on the consistency of learning. The noise transition matrix-based approaches (Kaneko, 2019) assume that there is a single transition probability between the noisy label and the ground-truth label, and this probability is independent of the individual samples. However, in real cases, the appearance of each sample has much influence on whether it can be misclassified. Another direction develops regularization techniques (Zhang, 2018), including explicit and implicit regularizations. This direction employs the regularization bias to overcome the label noise issue. Explicit regularization is added to the objective function.

Implicit regularization is designed for training algorithms. For instance, the confidence regularized self-training (CRST) (Zou, 2019) introduces the confidence regularization, treating the pseudo-labels as the continuous latent variables that are jointly optimized via alternating optimization. Nevertheless, both approaches introduce a permanent regularization bias, and the learned classier barely reaches the optimal performance. The additional supervision-based approaches (Lee, 2018) directly help the correction process. The main drawback of these approaches is that they require extra clean samples, making them expensive to apply in the largescale real-world scenarios. Although these approaches make some improvements, the revised segmented masks still contain somewhat noises that disturb the learning process and cause the deviation in the decision boundary.

3 OUR APPROACH

We propose an approach that combines the label correction and the sample reweighting based on the GAN framework to deal with the noisy labels challenge, and the detail is shown in Fig. 2. The generator (G) contains a multitask learning network call dual inference network that uses two U-Nets (Ronneberger, 2015) to jointly learn the regression from the image to the segmented mask and from the noisy label to the cleaned label. The upsampling parts in the U-Nets are related by the recurrent neural network. The visual cues in the different scales are all employed to guide the label correction analysis. The discriminator (D) measures the extent that revised label likely to be real, and the confidence score is employed to reweight the revised labels in the loss.

3.1 DUAL INFERENCE NETWORK

The noisy correction and segmentation are related tasks because the effectiveness of the data-driven method depends on the label quality. In addition, they both predict the segmented masks and can be implemented using an encoder-decoder framework. The image features that are obtained from the encoder are employed to guide the noisy correction process Li, 2019). However, the image features only work as a visual clue with a reduced dimension, providing neither the semantic nor the position information to benefit the label correction.

Therefore, we propose a dual inference network that simultaneously corrects the label and segments the image. Instead of combing the visual features and the label features in the embedded space for once, we argue that the feature maps in the neighboring scales relate to each other. Examples are illustrated in Fig. 3 that an image and the corresponding feature maps from different receptive fields are provided. The visual features and the mask features can be combined in the decoder part, because the feature maps in the deep layers are rich in localization information and high activation outputs. To explore the relation between the neighboring receptive fields, a recurrent neural network is employed as the multiscale analysis.

Figure 3: **Examples of the feature maps at scales 1, 2, 3 and 4 of the decoder (with 1 the lowest and 4 the highest resolution).**

Assume that W_k, H_k, D_k are width, height and channel number in the k-th scale feature maps $\mathbf{f}_k$ in the decoder. Two 3×3 residual learning blocks (He, 2016) with channel number D_k are applied at k-th scale to learn a latent $\mathbf{h}_k$, among which, one extracts the visual information at the same scale, and the other one propagates the latent information from the preceding scale.

$$\mathbf{h}_k = \mathbf{W}_{up}^k \mathbf{h}_{k-1} + \mathbf{W}_{mid}^k \mathbf{f}_k, \tag{1}$$

where $\mathbf{W}_{up}^k$ and $\mathbf{W}_{mid}^k$ are the network parameters to fulfill visual guidance and latent propagation. The order of convolution layer is BN-ReLU-Conv for information propagation.

In the training stage, compared to the global visual guidance approach (Li, 2019), our approach works on the visual and label information fusing through a recurrent method, so that the long-term dependency can be learned by iteratively combining the local context information from multiple scales. The whole process is fully data-driven and can be trained end-to-end. In the testing stage, only the image segmentation network is used for inference.

3.2 REWEIGHTING LOSS

After the label revision process, most of the labels are improved. However, some hard samples still contain noises due to the ambiguous factors hidden in the data. During the training process, the noisy labels might have a higher loss compared with the well-annotated ones. Different types and levels among the noises confuse the network and show a slower speed of loss descent.

Our goal is to construct a regression network to calculate the quality score for each sample and find the conflict information among the noisy labels. We introduce

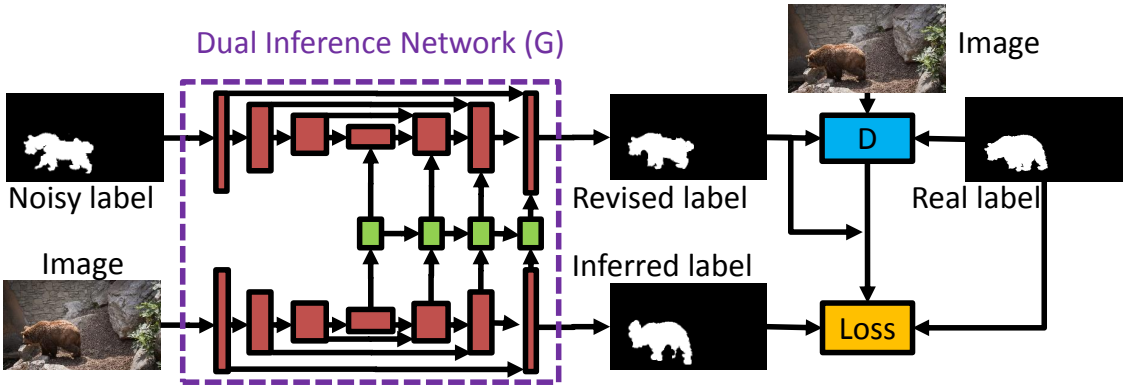

Figure 2: **The details of our approach. The dual inference network (G) simultaneously corrects the label and segments the image. The discriminator (D) judges the revised label and the score is used to reweight the revised labels.**

the GAN into the sample reweighting process. The revised label is generated by the dual inference network that works as the generator of the GAN, and the discriminator outputs 1 as the real label and 0 as the simulated label, regarding the output as the confidence of the real label.

The image and the revised/real label mask are concatenated as a 4 channels tensor. Our discriminator consists of seven 3×3 residual learning blocks with the stride of 2 and one fully-connected layer. Our generator is trained using both a pixel-wise binary cross-entropy (BCE) loss and an adversarial loss. The BCE loss ensures a high local similarity between a refined label map and the corresponding ground truth. The adversarial loss tries to guarantee the refined label maps residing in the manifold of the ground truth.

Assume the i-th image is represented by $\mathbf{x}_i$, its original noisy label mask is $\mathbf{y}_i$, the refined label mask that is obtained by the dual inference network is $\mathbf{z}_i = G(\mathbf{x}_i, \mathbf{y}_i)$, and the real label mask is $\mathbf{t}_i$. The adversarial loss is constructed as follows

$$L_{adv} = E_{\mathbf{x}_i,\mathbf{t}_i}[logD(\mathbf{x}_i, \mathbf{t}_i)] + E_{\mathbf{x}_i,\mathbf{z}_i}[log(1-D(\mathbf{x}_i, \mathbf{z}_i))]. \tag{2}$$

The BCE loss between a high-quality label map $\mathbf{t}_i$ and the inferred map $\mathbf{r}_i$ is calculated as

$$L_{bce} = E_{\mathbf{r}_i,\mathbf{t}_i}[-\frac{1}{M}\sum_j (t_{ij} log\mathbf{r}_{ij} + (1-t_{ij}) log(1-\mathbf{r}_{ij}))], \tag{3}$$

where M is the number of pixels in a map and j is the index of each pixel. The final objective is then formulated as

$$L_{total} = \min_G(\max_D L_{adv}(G, D) + \lambda L_{bce}(G)), \tag{4}$$

where λ controls the importance of two losses, and the selection of λ is explained in the ablation study. In the training stage, the segmentation network is firstly learned with the predicted masks and the correctly annotated samples by optimizing the L_{BCE} loss. Then, the predicted segmented mask works as the noisy label $\mathbf{y}_i$, and cooperates with the ground truth t_i to learn the label correction network by optimizing the L_{BCE} loss. After that, the discriminator is learned using the L_{adv} loss and the revised and real label mask pairs. Finally, the generator is finetuned using the L_{total} loss and the noisy and revised label mask pairs.

The discriminator outputs the confidence that is related to the label trueness. Therefore, the revised masks tending to be the ground truth receive large weights while the still noisy labels obtain small factors in the weighted binary cross-entropy (WBCE) loss function.

$$L_{wbce} = -\sum_i \mathbf{w}_i \sum_j (t_{ij} log\mathbf{r}_{ij} + (1-t_{ij}) log(1-\mathbf{r}_{ij})), \tag{5}$$

where the weight $\mathbf{w}_i$ is the label trueness score that is obtained from the discriminator. The WBCE loss is used to finetune the segmentation network. The revised labels, although they may contain small portion of error segmented pixels, the remaining correct labeled pixels take part in the learning process with relatively reasonable contribution.

4 EXPERIMENTAL RESULTS

In this section, we compare the performance of our proposed approach to other approaches handling the noisy labels.

4.1 DATASETS

We verify our proposed approach on the publicly available PASCAL VOC 2012 dataset (Everingham, 2015) and Shining3D dental dataset (Tian, 2019c).

The PASCAL VOC 2012 dataset has 20 object categories and one background category. It was split into a training set of 1,464 images, a validation set of 1,449 images and a test set of 1,456 images. We reannotated 20% of the samples (including the training and validation sets), and if the annotated samples and the ground truth had intersection-over-union (IoU) less than 0.95, the new annotated samples were regarded as the noisy labels. Following the common practice, we increased the number of training images to 10,582 by augmentation. The performance of our method and other state-of-the-art methods were evaluated on the validation set and test set.

The Shining3D was a set of 47 videos of human mouths generated by a 3D dental scanning device that is often used for research. Each video came from a hospital patient who was selected randomly. We made some changes to ensure that privacy would be maintained. We randomly selected and annotated 7,800 images from these videos consisting of a training set of 5,800 images from 40 people and a validation set of 2,000 images from the remaining 7 people. The image sizes were fixed to be 640 pixels in width and 480 pixels in height. Four researchers from our university annotated the training and validation images; another 4 researchers reannotated the same images to ensure correctness. When the labels disagreed, we employed another person for evaluation, then we obtain the noisy labels and the ground truth. Each annotator used a software named LabelMe to mark the boundary and classify each region as tooth, gum, jaw, lip, cheek, or other soft tissues. For application purposes, we set tooth and gum as classes of interest and all other soft tissues as classes that were less relevant.

4.2 EVALUATION CRITERIA

We used an evaluation criteria that others have used in the published research to compare our work to state-of-the-art approaches. The class-wise IoU between the ground truth mask and the predicted mask was employed to measure region-based segmentation similarity. Specifically, given a predicted mask P and corresponding ground truth mask G, IoU was defined as $IoU = \frac{P \cap G}{P \cup G}$.

4.3 IMPLEMENTATION DETAILS

We used a workstation with an Intel i7-4790 3.6 GHz CPU, 32GB memory, and NVIDIA GTX Titan X graphics. Our algorithm to verify performance was based on Tensorflow.

The images in the PASCAL VOC 2012 dataset and Shining3D dental dataset were cropped and resized to the size of 513×513 and 480×480, respectively. In our implementation, the backbone was U-Net (Ronneberger, 2015). We followed the practice (Zhang, 2019) to use the weights pretrained on the ImageNet (Deng, 2009) to initialize the backbone network. All weights of newly added layers were initialized with Gaussian distribution of variance 0.01 and mean 0. The parameter λ was chosen according to the experimental results, and the details were described in the ablation study. The GAN network was trained on stochastic gradient descent (SGD) with a momentum of 0.9 and a weight decay of 0.0005. To fairly compare with the approach (Jiang, 2018), we also used the "poly" policy to adjust the learning rate, and the initial learning rate was 0.007. Batchsize was set to 16 for the PASCAL VOC 2012 dataset and 20 for the Shining3D dental dataset.

4.4 ABLATION STUDY

We perform extensive ablation studies to observe the effects of several important components of our approach. These experiments are performed on the PASCAL VOC 2012 dataset only.

U-Net Structure. Schemes with different number of the scales in the U-Net are evaluated, and the IoU comparisons are illustrated in Fig. 4. In our experiment, $n = 4$ scales achieve the competing effectiveness. If more scale are used in the scheme, the nonlinear mapping between the image and the segmented mask is modelled better. However, simply increasing the complexity of the network may cause overfitting. Therefore, we employ an U-Net with 4 scales for effectiveness comparison in the following experiments.

Loss Function. The parameter λ is introduced to control the relative weights of the binary cross-entropy loss and the adversarial loss. We conduct experiments by setting different weight λ of the binary cross-entropy loss and show the results in Fig. 5. Small λ will deteriorate the learning process, while the noisy labels are not improved if the weight λ is too big. Only suitable λ would take effect at the training stage. The weight $\lambda = 10^{-4}$ yields the best performance, which outperforms the runner-up with an improvement of IoU of 0.2.

Effectiveness. Finally, we evaluate the effectiveness of the label correction and sample reweighting in image segmentation. The IoU comparison can be seen in Tab. 1. When the label correction is employed, the IoU is increased to 85.4, and the sample reweighting can further improve the error by a margin of approximately 0.7.

Figure 6: **Experiment results on the PASCAL VOC 2012 validation dataset. The 1st (top) row is the input images, and the 2nd-4th rows are the predicted segmentation masks of the IGAN and our approach, and the corresponding ground truth. (a) Accurate segmentation results. (b) Inaccurate segmentation results.**

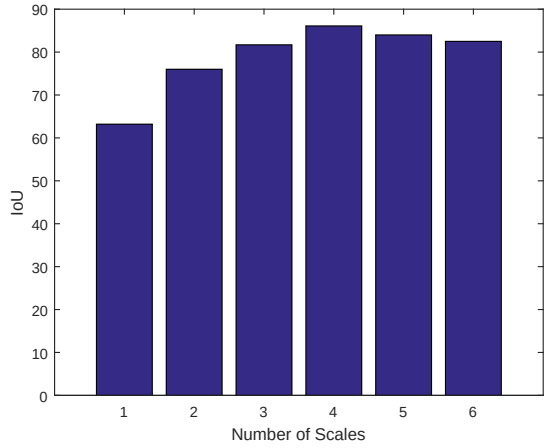

Figure 4: **Evaluation of the number of scales in the U-Net.**

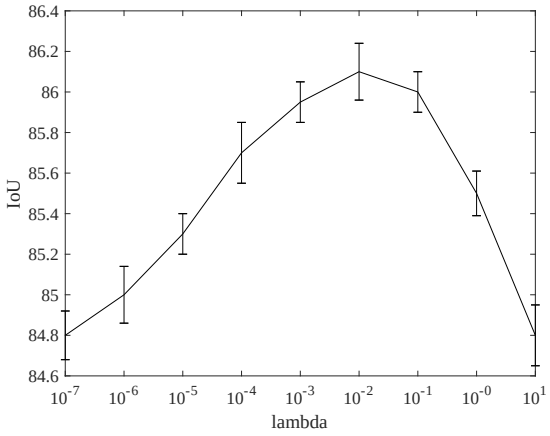

Figure 5: **Quantitative analysis of weight λ.**

Table 1: **Experimental results on the PASCAL VOC 2012 dataset. The IoU improvements of each module are reported.**

Method	Noisy label	Revised label	+Reweight
IoU	84.3	85.4	86.1

4.5 EVALUATION ON THE PASCAL VOC 2012 DATASET

We compare our approach to other approaches dealing with the noisy labels in image segmentation, with experimental results shown in Tab. 2.

Table 2: **Effectiveness comparison on the PASCAL VOC 2012 test set and the Shining3D dental dataset. * means the corresponding approach is reimplemented.**

Approach	VOC12	Shining3D
	IoU(%)	IoU(%)
DifNet (Jiang, 2018)	72.7	79.6
PixelDCL (Gao, 2017)	73.3	80.5
SPN (Liu, 2017)	79.2	86.2
AAF (Ke, 2018)*	81.6	88.4
PSPNet (Zhao, 2017)	84.9	91.5
LLGC (Li, 2019)*	79.1	85.5
LCN (Li, 2019)*	84.2	91.4
IGAN (Yang, 2019)*	85.0	92.0
Ours	86.1	92.9

We firstly compare with different context exploitation approaches. The DifNet uses a method that is similar with the cascade of random walks to propagate the local information, obtaining only IoU of 72.7. The pixel decon-

volutional layer (PixelDCL) and the spatial propagation networks (SPN) employ the specially designed subpixel convolution and directional convolution to learn the local relation, and receive the IoU of 73.3 and 79.2, respectively. The adaptive affinity field (AAF) combines multiple convolution using the adversarial weighting, achieving IoU of 81.6. The pyramid scene parsing network (PSPNet) uses the different-region based context aggregation to generate the global context, and uses this global context to guide the inference process, which works effectively in segmentation.

Then, we evaluate the approaches dealing with the noisy labels. The 'LLGC' is an unsupervised approach, using the low-level features of the superpixels for local and global consistency learning. The correction results are sensitive to the intervariance due to the lack of the supervision information, receiving IoU of 79.1. The label cleaning net (LCN) employs a small network containing the multiscale analysis to directly predict the segmented masks, obtaining an improvement of IoU of 5.1 when compared with the unsupervised approach. The 'IGAN' also uses the U-Net to revise the label mask. However, neither the layer relation nor the sample reweighting is employed, as a result, it only gets the IoU of 85.0. Based on the IGAN, our method combines a recurrent network to learn the long-term dependency and introduces the discrimination score to the guide the sample reweighting, receiving an additional IoU improvement of 1.1.

The segmentation results of the IGAN and our approach on the PASCAL VOC 2012 validation dataset are shown in Fig. 6. Our approach is consistently accurate in the output masks (some misclassified pixels in IGAN are now correctly classified), and the details and object boundaries are clear with the help of the context exploitation module, such as the sheep and cow images in Fig. 6(a).

However, there are still some ambiguous results, for example, the tail of the motorcycle image in Fig. 6(b). Inexact segmentation of the details partially due to missing global information in the U-Net structure, and the global semantic knowledge such as the object attributes with the reduced size can be employed to combine with the details to improve the prediction result.

4.6 EVALUATION ON THE SHINING3D DATASET

In this section, we discuss our approach compared to other approaches on the Shining3D dental dataset. Effectiveness comparisons are provided in Tab. 2. Note that our results on Shining3D dental dataset give conclusions similar to those from our results on the PASCAL VOC 2012 dataset.

Some of our label correction results on the Shining3D dental dataset are shown in Fig. 7. The 1st (top) row is the input images, and the 2nd-4th rows are the noisy label, the revised label, and the real label, respectively. Successful label correction results in Fig. 7(a) demonstrate that our method is robust to variations in dental shape, camera motion, and background clutter. Although the cheek has high similarity with the gum, the context exploitation enhances the discriminant capability to distinguish the gum region from the cheek region. The tooth and white tongue guard are also ambiguous, and our label correction module identifies the tongue guard to generate the accurate tooth region. Fig. 7(b) shows some fail revision results. The dental calculus is originally labeled as the background, and the label correction network only rescues a small part of the tooth region, partially due to the scarcity of the similar samples in the training data.

Fig. 8 shows the segmentation results. Fig. 8(a) illustrates the accurate prediction masks. The effectiveness of the segmentation network is improved when the training samples are revised and reweighted by the generator and the discriminator. Fig. 8(b) illustrates the inaccurate segmented results. Notice that the shadow that is brought by the hardware and the tooth, greatly affects the segmented result, due to the deviation from typical characteristics of the appearance. Other factors such as the tooth variation and the gum dirt enlarge the intra-variance, and soft tissues, for example, cheek and tongue, have almost the same appearance as the gum, which makes the intervariance to be small.

In the future, we will extend our method to handle the situations in inaccurate results by employing the hard sample mining strategy. The processing time for each 480×480 image in the Shining3D dental dataset is approximately 74 milliseconds, and we will also work on methods to increase the efficiency so that our method is practical to use for the real-time applications such as the autonomous driving.

5 CONCLUSION

Our experimental results show that the GAN-based approach is an effective approach to handle the noisy label masks by simultaneously label correction and sample reweighting. Specifically, we present a dual inference network to recurrently combine visual clues from different receptive fields, and introduce the discriminant score to adjust the sample weight in the loss function. Although our method may arrive at inexact segmentation due to factors such as shadows, our method is robust to variations in object shape, camera motion, and back-

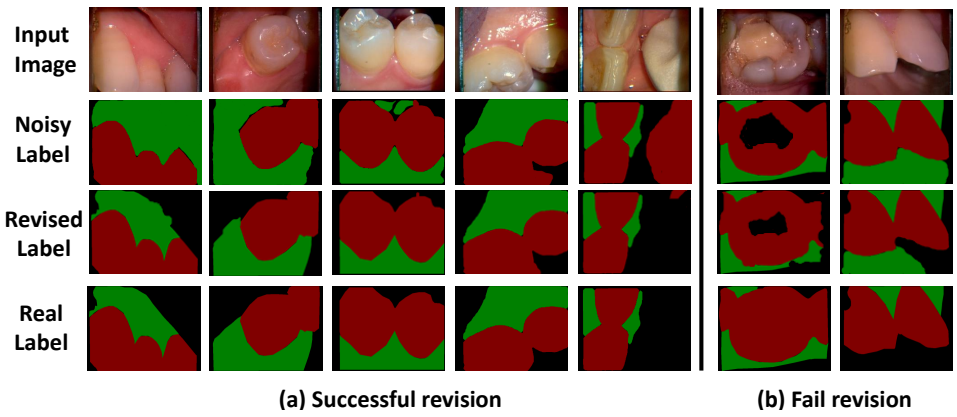

Figure 7: **Label correction results on the Shining3D dental dataset. The 1st (top) row is the input images, and the 2nd-4th rows are the noisy label, revised label, and real label, respectively. (a) Successful label correction results. (b) Fail label correction results.**

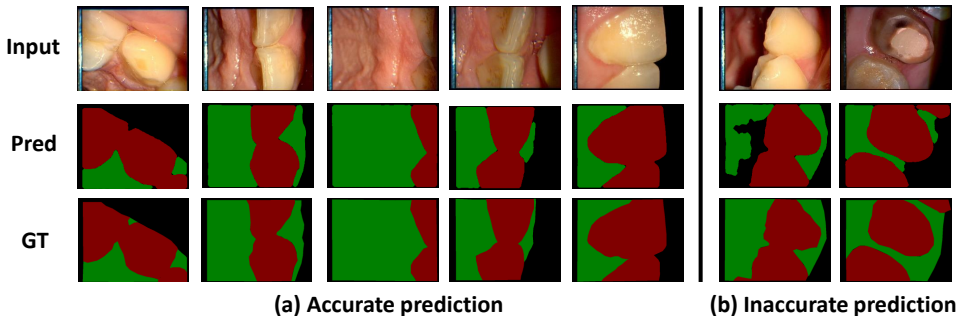

Figure 8: **Segmentation result on the Shining3D dental dataset. The 1st (top) row is the input images, and the 2nd-3rd rows are the predicted segmentation masks and the corresponding ground truth. (a) Accurate segmentation results. (b) Inaccurate segmentation results.**

ground clutter.

Acknowledgements

This work was supported by the National Key R&D Program of China under grant 2016YFB1000400, the National Natural Science Foundation of China under Grant 61972351, the Scientific Research Foundation of National Health and Family Planning Commission under grant WKJ-ZJ-1814, the Key R&D Plan of Zhejiang Province under grant 2019C03002, the Natural Science Foundation of Zhejiang Province under Grant LY19F030005, and the Hangzhou Major Science and Technology Innovation Project under grant 20172011A038.

References

B. Tu, CH-L. Zhou, D-B. He, S-Y. Huang, and P. Antonio (2020). Hyperspectral Classification With Noisy Label Detection via Superpixel-to-Pixel Weighting Distance.*IEEE Transactions on Geoscience and Remote Sensing*.**111**(1):1-16.

E. Mark, E. SM Ali, V.G. Luc, W. C. KI, W. John, and Z. Andrew (2015). The pascal visual object classes challenge: A retrospective.*International journal of computer vision***111**(1):98-136.

G-R. Wang, P. Luo, L. Lin, and X-Y. Wang (2017). Learning object interactions and descriptions for semantic image segmentation.*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5859-5867.

H. Lu, Y-T. Dai, CH-H. Shen, and S-C. Xu (2019). Indices matter: Learning to index for deep image matting.*Proceedings of the IEEE International Conference on Computer Vision*, 3266-3275.

H-Y. Gao, H. Yuan, Z-Y. Wang, and SH-W. Ji (2017). Pixel deconvolutional networks.*arXiv preprint arXiv:1705.06820*.

H-SH. Zhao, J-P. Shi, X-J. Qi, X-G. Wang, and J-Y. Ji-

a (2017). Pyramid scene parsing network.*Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881-2890.

J. Deng, W. Dong, S. Richard, L-J. Li, K. Li, and F-F. Li (2009). Imagenet: A large-scale hierarchical image database.*2009 IEEE conference on computer vision and pattern recognition*, 248-255.

K-M. He, X-Y. Hang, SH-Q. Ren, and J. Sun (2016). Deep residual learning for image recognition.*Proceedings of the IEEE conference on computer vision and pattern recognition*, 770-778.

K. Takuhiro, U. Yoshitaka,and H. Tatsuya (2019). Label-noise robust generative adversarial networks.*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 246-2476.

K. Tsung. Wei, H. Wang, Jyh. Jing, ZW. Liu, Yu, and Y. Stella. X (2018). Adaptive affinity fields for semantic segmentation.*Proceedings of the European Conference on Computer Vision*, 587-602.

L-C. Chen , Y-K. Zhu, P. George, S. Florian, and A. Hartwig (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation.*Proceedings of the European conference on computer vision*, 801-818.

L. Kuang. Huei, X-D. He, L. Zhang, and L-J. Yang (2018). Cleannet: Transfer learning for scalable image classifier training with label noise.*Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5447-5456.

N. Hyeonwoo, H. Seunghoon, and H. Bohyung (2015). Learning deconvolution network for semantic segmentation.*Proceedings of the IEEE international conference on computer vision*, 1520-1528.

O. Marin, K. Ivan, B. Petra,and S. Sinisa (2019). In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images.*Proceedings of the IEEE conference on computer vision and pattern recognition*, 12607-12616.

P. Jiang, F-L. Gu, Y-H. Wang, CH-H. Tu, and B-Q. Chen (2018). Difnet: Semantic segmentation by diffusion networks.*Advances in Neural Information Processing Systems*, 1630-1639.

P-P. Zhang, W. Liu, H-Y. Wang, Y-J. Lei, and H-CH. Lu (2019). Deep gated attention networks for large-scale street-level scene segmentation.*Pattern Recognition***88**: 702-714.

R. Olaf, F. Philipp, and B. Thomas (2015). U-net: Convolutional networks for biomedical image segmentation.*International Conference on Medical image computing and computer-assisted intervention*, 234-241.

S-F. Liu, D. Mello. Shalini, J-W. Gu, G-Y. Zhong, M-H. Yang, and K. Jan (2017). Learning affinity via spatial propagation networks.*Advances in Neural Information Processing Systems*, 1520-1530.

S. Lin, Y-W. Li, Z-M. Li, G. Yu, H-B. Sun, J. Sun, and N-N. Zheng (2019). Learnable Tree Filter for Structure-preserving Feature Transform.*Advances in Neural Information Processing Systems*, 1709-1719.

T. Agnieszka, N. Nassir, and A. Shadi (2019). Learn to estimate labels uncertainty for quality assurance.*Tomczack, Agnieszka and Navab, Nassir and Albarqouni, Shadi*,arXiv preprint arXiv:1909.08058.

T. Shusuke, K. Yusuke, M. Yusuke, A. Hiroyuki, F. Masashi, Y. Akihiko, K. Masanobu, and H. Tatsuya (2019). Multi-Stage Pathological Image Classification using Semantic Segmentation.*Proceedings of the IEEE International Conference on Computer Vision*, 10702-10711.

W-G. Wang, J-B. Shen, and H-B. Ling (2018). A deep network solution for attention and aesthetics aware photo cropping.*IEEE transactions on pattern analysis and machine intelligence*, 1531-1544.

Y-L. Li, L-H. Jia, Z-D. Wang, Y. Qian, and H. Qiao (2019). Un-supervised and semi-supervised hand segmentation in egocentric images with noisy label learning.*Neurocomputing* **334**: 11-24.

Y-Q. Yang, ZH-W. Wang, J-G. Liu, Ch. Kwang. Ting, and X. Yang (2019). Label Refinement with an Iterative Generative Adversarial Network for Boosting Retinal Vessel Segmentation.*arXiv preprint arXiv:1912.02589*.

Y. Tian, H-Y. Wang,and X. Wang (2017). Object localization via evaluation multi-task learning.*Neurocomputing***253**: 34-41.

Y. Tian, G. Judith, X. Wang, W-G. Chen, J-X. Gao, Y-J. Zhang, and X-L. Li (2018). Lane marking detection via deep convolutional neural network.*Neurocomputing***280**: 46-55.

Y. Tian, W. Hu, H-S. Jiang, and J-CH. Wu (2019a). Densely connected attentional pyramid residual network for human pose estimation.*Neurocomputing***347**: 13-23.

Y. Tian, G. Judith, X. Wang, J-Y. Li, and Y-ZH. Yu (2019b). Traffic sign detection using a multi-scale recurrent attention network.*IEEE Transactions on Intelligent Transportation Systems***20**(12):4466-4475.

Y. Tian, G-H. Cheng, G. Judith,SH-H. Yu, CH. Song, and B-L. Yang (2019c). Joint temporal context exploita-

tion and active learning for video segmentation.*Pattern Recognition*, 107-118.

Y. Tian, X. Wang, J-CH. Wu, R-L. Wang, and B-L. Yang (2019d). Multi-scale hierarchical residual network for dense captioning.*Journal of Artificial Intelligence Research***64**: 181-196.

Y. Zou, ZH-D. Yu, X-F. Liu, K. BVK, and J-S. Wang (2019). Confidence regularized self-trainin.*Proceedings of the IEEE International Conference on Computer Vision*, 5982-5991.

ZH-L. Zhang, and S. Mert (2018). Generalized cross entropy loss for training deep neural networks with noisy labels.*Advances in neural information processing systems*, 8778–8788.