An Efficient Weakly-Supervised Learning Method for Optic Disc Segmentation

Yang Wen[1], Leiting Chen[1,2], Lifeng Qiao[3,4], Chuan Zhou*[1], Shuo Xi[1], Rui Guo[1], Yu Deng[5]
[1]Key Laboratory of Digital Media Technology of Sichuan Province, School of Computer Science and Engineering
[2]Institute of Electronic and Information Engineering in Guangdong, [3]School of Medicine
University of Electronic Science and Technology of China
[4]Ophthalmology Department of Sichuan Provincial People's Hospital
[5]Dept. Biomedical Engineering, King's College London, London, UK
young.wen@foxmail.com, richardchen@uestc.edu.cn, tgqiao2004@163.com, zhouchuan@uestc.edu.cn
{xishuo,guorui9102}@std.uestc.edu.cn, malikteng@foxmail.com

***Abstract*—Accurate optic disc segmentation plays an essential role in the early diagnosis of glaucoma, which has been a major cause of irreversible blindness for the past decade. Recently, U-shape Convolutional Neural Network (CNN) models have achieved favourable performance in optic disc segmentation. However, it is worth noting that these models require a large number of pixel-level annotations while these annotations are difficult to obtain in clinical practice. As a solution, weakly-supervised training methods are commonly implemented, but it will provoke U-shape CNN generating inaccurate, diluted, and grid-like segmentation results. In this paper, we propose a novel Hybrid Network (HyNet) to solve the issue above. HyNet consists of a U-shape backbone hybridized with a cross-scale connection structure, which makes better use of multi-scale visual semantics. Nevertheless, the generalization ability of HyNet is affected by the domain shift among different datasets. Therefore, we innovatively combine weakly- and fully-supervised training methods, namely Hybrid Process (HyProcess), to solve the domain shift problem. Experimental results on ONHSD, DRIONS-DB, and DRISHTI-GS datasets show that our model outperforms the state-of-the-art, reaching Dice of 82.39(%), 93.72(%), and 95.34(%) respectively. Additionally, our ablation study validates the effectiveness of HyNet along with HyProcess, and further analysis reflects their value in clinical practice.**

***Index Terms*—Segmentation, Optic disc, Convolutional neural network, Weakly supervision.**

I. Introduction

Glaucoma, an incurable disease, has been the second leading cause of irreversible blindness within the past decade [1]. Automatic optic disc (OD) segmentation in funduscopic images has been gaining constant attention from global researchers for early-stage diagnosis and deterioration control of glaucoma [2]–[6].

In the past few years, deep convolutional neural networks with U-shape structure have shown their effectiveness in accurate OD segmentation [5], [7]–[9]. However, there are still two critical shortcomings that hinder their application in clinical practice. First, conventional fully-supervised training protocols demand extensive pixel-level annotations, which are usually unavailable due to consuming of time and requirements of professional skills. Second, since different scanners could bring significant variance in observation angles, color distribution, and frame scale of images, causing the domain shift problem that significantly jeopardizes the generalization ability of deep neural networks [10]. As a result, a well-trained model may produce poor segmentation when switched to a new type of image (see Fig. 1).

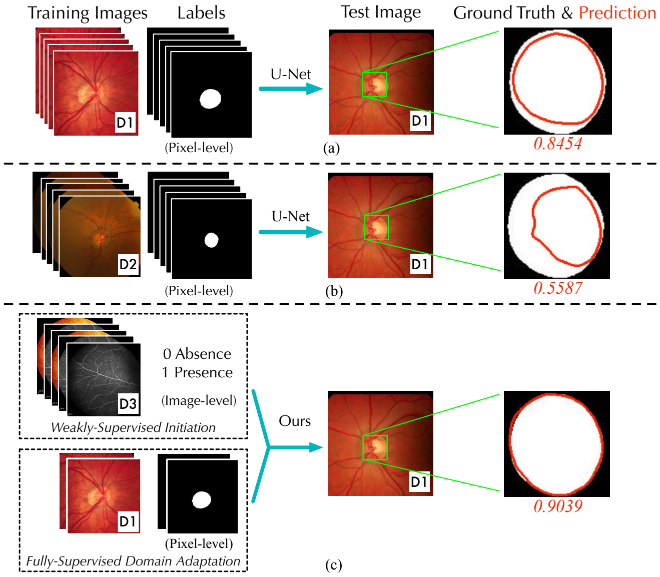


Fig. 1. Segmentation degradation caused by domain shift. Domain **D1** stands for DRIONS-DB dataset, domain **D2** stands for DRISHTI-GS dataset, and domain **D3** stands for private SPPH dataset. Both (a) and (b) are trained with numerous pixel-level annotations, while (c) is trained with image-level labels for initiation and later finetuned with a few pixel-level annotations. As seen in (b), the Jaccard degrades from 84.54(%) to 55.87(%), when U-Net trained on **D2** and tested on **D1**. Comparatively, our method achieves the best results on target domain **D1** with less reliance on pixel-level annotations.

A variety of studies have been conducted to address the above two issues. Firstly, to reduce the reliance on pixel-level annotations, weakly-supervised segmentation methods

*Chuan Zhou is the corresponding author.

[11]–[13] emerge as a competitive alternative because they only require image-level annotations. The image-level labels, which indicate the absence and presence of targets, are much easier to obtain. Secondly, an adversarial learning architecture is recently introduced to solve the domain shift problem of OD segmentation [10]. However, they still suffer from some shortcomings, namely inaccurate segmentation due to weak and inadequate supervision, and the adversarial models still require a large number of pixel-level annotated samples for training. In terms of clinical use, accuracy, cost-effectiveness and flexibility are all critical to deep neural networks. Thus, a method that could simultaneously reduce the reliance on pixel-level annotations, overcome the domain shift and obtain accurate segmentation is highly desired.

In this work, we aim to jointly address the aforementioned issues of the deep neural networks in order to establish a practical solution for clinical OD segmentation. Specifically, an improved U-shape CNN with cross-scale connection, namely Hybrid Network (HyNet), and a Hybrid Process (HyProcess) training method, are proposed. In HyNet, unlike the previous models [8], of which residual connection is applied to either encoding or decoding path, we innovatively craft direct connections onto both up- and down-stream of the network. The HyProcess scheme is designed to train the HyNet under both weakly- and fully-supervised manner. Through weakly-supervised training with image-level annotations, our model is initiated with abundant localization and morphological information of the OD in a more economical way. Afterwards, fully-supervised domain adaptation is crafted with a small number of pixel-level annotations, introducing both domain characteristics and pixel-level details to overcome the domain shift and fulfill inadequate supervision. Extensive experiments on three public datasets demonstrate the effectiveness and efficiency of our methods.

Our main contributions are as below:

- We design a novel HyProcess scheme for OD segmentation, which jointly utilizes weakly- and fully-supervised learning to reduce the expense and enhance the robustness simultaneously in the funduscopic domain.
- We propose a HyNet architecture embedded with a novel cross-scale connection structure, which helps capture multi-level visual semantics, preserve spatial details and still maintain high efficiency.
- Experiments are conducted on three public OD segmentation datasets (DRIONS-DB, DRISHTI-GS, and ONHSD), and the experimental results of our method outperform the state-of-the-art, suggesting that our method can provide accurate, robust, and economical solutions for clinical practice.

II. Related Work

A. Optic Disc Segmentation

Studies on OD segmentation using funduscopic images has been ongoing for many years. Early work employed the hand-crafted visual features from Hough transformation [2], saliency [14], morphological operation [3] and stereo image pairs [4] to obtain segmentation results. Recently, the U-shape convolutional neural networks (CNNs) proposed by [15] greatly boost the performance on OD segmentation thanks to their ability in capturing both high-level visual semantics and low-level characteristics. Several variants are proposed for further enhancement in couple with densely connection [5], context extractor [9], deep supervision [7], and generative adversarial networks [6]. Though promising are these models, they demand a large number of expensive pixel-level annotations for training and suffer performance degradation from domain shift over different datasets. Our HyProcess helps address this domain shift challenge and reduces reliance on costly annotations.

B. Domain Adaptation

In the last few years, domain adaptation methods on medical image analysis have received increasing attention due to their improved generalization capabilities. Among all approaches, generative adversarial networks (GANs) achieve the most promising results by either applying latent feature alignment [16] or crafting domain transformation on input images [17]. Recently, [10] presented an output space adversarial learning framework to overcome the domain shift specifically for OD segmentation. However, a large number of pixel-level annotations are still necessary for GANs, and the stability of adversarial learning is worrisome [18], [19].

C. Weakly-Supervised Segmentation with Medical Images

Studies on weakly-supervised segmentation are mostly conducted on general color images [20], [21], but rarely on medical images. [12] proposed a multiple instance learning framework with both weakly-supervised learning and clustering for histopathology image segmentation. [13] developed a weakly-supervised fully convolutional networks and multi-scale fusion for cancerous regions segmentation in histopathology images. [11] proposed a CNN with global average pooling layers for pulmonary segmentation. Different from previous studies on weakly-supervised OD segmentation [22], [23], we validate the ability of weakly-supervised approaches on tackling the domain shift problem, improve the robustness of CNN model, and achieve a promising generalization performance on challenging OD segmentation datasets.

III. Methods

In this section, we present the Hybrid Network (HyNet) and the Hybrid Process (HyProcess) to address the problem of OD segmentation. The overall pipeline is displayed in Fig. 2. The HyNet consists of two key components: U-shape backbone and cross-scale connection, while the HyProcess includes weakly-supervised initiation and fully-supervised domain adaptation. Different from the previous works that require preparatory ROI extraction of the OD region [9], [10], we conduct OD segmentation directly on the full-sized image.

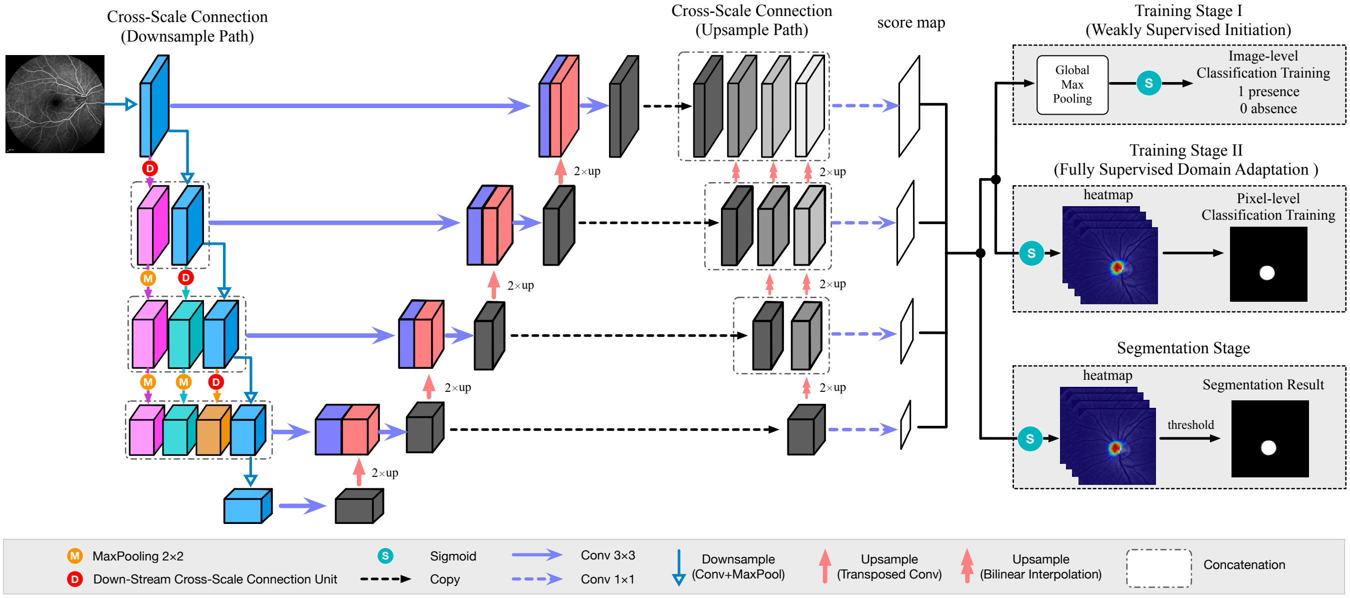


Fig. 2. The overall pipeline of our method. HyNet, an improved U-shape architecture equipped with the cross-scale connection structure, is illustrated on the left part and HyProcess is shown on the right part.

A. *HyNet*

1) U-shape backbone: We build our architecture based on the U-shape convolutional neural networks, and design a top-down encoding path and a bottom-up decoding path [15]. In this architecture, input images are convolved and downsampled layer by layer in the encoding path, therein generate feature maps with different scales. Later, transposed convolutional layers are used in the decoding path to restore the sizes of feature maps. Finally, both high-level and low-level visual information will be retained by fusing the encoding and decoding results.

2) Cross-Scale Connection: Due to the failure of capturing detailed visual characteristics from inadequate supervision in weakly-supervised training, models with original U-shape structure (M-Net and BRU-Net in Fig. 5) produce fuzzy results. Thus, a cross-scale connection (CSC) structure is proposed to address this problem. Different from the previous method [8], our CSC takes both down-stream and up-stream paths into consideration, which leads to better capability in the fusion of multi-scale features.

In the downsample path of HyNet, we utilize a combination of pooling layers and dilated convolutional filters to extract information-rich visual features and bring them among different layers (see Fig. 3). It first generates multi-scale feature maps, which will be subsequently added element-wisely and later aggregated by a 1×1 convolutional layer. After being downsampled by the max-pooling layers, feature maps across different scales are concatenated correspondingly as the final encoding outputs. Unlike the usual ways of preserving spatial semantics, such as ASPP structure [24], [25] and pooling-based method [26], we use such a consolidated method to balance the performance and efficiency. On the one hand, although the dilated convolutional operations enlarge the receptive fields, they bring lots of extra computational cost. On the other hand, pooling operations can capture local features without extra learning parameters, but inevitably lose spatial details. Thus, we design the down-stream block to maximize their strengths while avoiding weaknesses. As seen in experiments, the proposed method achieves improved performance and still be highly efficient.

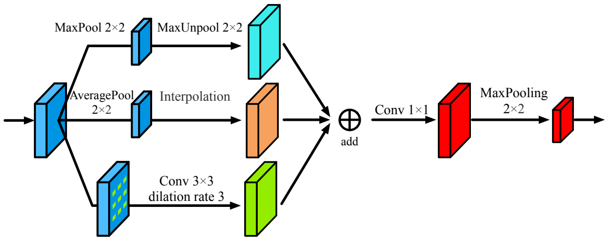


Fig. 3. Detail structure of our down-stream CSC unit.

In the upsample path of HyNet, the bilinear interpolation and concatenation are utilized (up-stream part of Fig. 2). Firstly, bilinear interpolation is applied to encoding outputs, restoring them to bigger sizes. Then, concatenation of the size-restored feature maps and the same-level decoding outputs are subsequently aggregated by a 1×1 convolutional layer. The reason for using bilinear interpolation instead of transposed convolution is to avoid the extra computational cost and the grid-like results [27] as shown in Fig. 5e and Fig. 5g.

Additionally, we introduce deep supervision [7], [28]–[30] to our architecture. After the upsample path of CSC, a total of four score maps are generated through 1×1 convolutional layer. Later, all four score maps will be used to calculate the loss for backpropagation within both weakly- and fully-supervised training stages.

B. HyProcess

We propose a novel HyProcess scheme, which is a two-stage hybrid-supervised training protocol, for OD segmentation. In the first stage, weakly-supervised training is employed to roughly locate and outline the OD, namely *Weakly-Supervised Initiation*. In the second stage, fully-supervised training is used for capturing detailed visual semantics and domain-specific characteristics with limited pixel-level annotated samples to overcome the domain shift problem, namely *Fully-Supervised Domain Adaptation*.

1) Weakly-Supervised Initiation: A weakly-supervised learning process is used to initiate the model with image-level annotations. Similar to [20], a global max-pooling layer is used on the final score map to aggregate spatial information into a global prediction that indicates the probability of either the image possesses the OD or not. In this way, the model will recognize the primary characteristics of OD.

2) Fully-Supervised Domain Adaptation: After the weakly-supervised initiation, the fully-supervised domain adaptation is used for further improvement. More specifically, it is a pixel-level classification training with images in the target domain. Compared to the conventional fully-supervised training, our approach only requires a few samples because the majority of information and visual characteristics of the OD have been captured at the weakly-supervised stage. Eventually, a considerable improvement of segmentation is obtained, and the domain shift is overcome, as shown in Fig. 7.

We use standard binary cross entropy and dice coefficient as optimization targets (*i.e.,* loss functions) for weakly- and fully-supervised learning, respectively. The cross entropy is defined as:

$$cross\ entropy = -\sum t(x)\log(p(x)) \tag{1}$$

where p and t corresponding to ”prediction” and ”target”. And the dice coefficient loss is defined as:

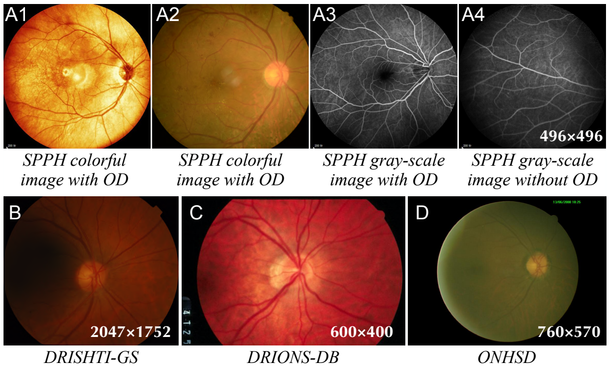
$$dice\ coefficient = 1 - \frac{2\sum_{i\in S} p_i \cdot y_i}{\sum_{i\in S} p_i^2 + \sum_{i\in S} y_i^2} \tag{2}$$

where S is the collection of all pixels in the image; p is predicted probability map; y is binary ground truth mask. For weakly-supervised learning, we only use the cross entropy loss. As for fully-supervised learning, we use the weighted sum of cross entropy and dice coefficient as the final segmentation loss.

IV. Experiments

A. Experimental Setup

1) Dataset: To assess the performance of the proposed models, we conducted experiments on three public OD segmentation datasets, that are DRIONS-DB [31], DRISHTI-GS [32], and ONHSD [33]. Additionally, we collected another private SPPH retinal fundus dataset, containing 7,351 grayscale images and 119 color images out of 173 patients, from a regional hospital that serves more than one million people annually. Different from the three public datasets, of which all images possess the ODs, only 1,905 images of SPPH possess the ODs while the other 5,565 images do not (see group A of Fig. 4). All images have been labelled with the absence and presence of the OD by experienced ophthalmologists. The statistics of these four datasets are listed in Table I and their appearances in different domains are shown in Fig. 4. The whole SPPH dataset is only used for weakly-supervised initiation, while the other three public datasets are used for fully-supervised domain adaptation. Moreover, we first randomly select 50 images from each public dataset, some of which are used for training and others for validation. Then, the rest images are used as the testing set.



SPPH colorful image with OD *SPPH colorful image with OD* *SPPH gray-scale image with OD* *SPPH gray-scale image without OD*

DRISHTI-GS *DRIONS-DB* *ONHSD*

Fig. 4. Comparison of fundus images from different datasets. Note that they have very different color distributions and frame sizes.

TABLE I
Summarization of the datasets. Here C denotes color image and G denotes grayscale image, respectively.

Dataset	# of samples	Image size	Type
SPPH	Total: 7470	496 × 496	C + G
DRISHTI-GS	Train:50, Test:51	2047 × 1752	C
DRIONS-DB	Total:110	600 × 400	C
ONHSD	Total:99	760 × 570	C

2) Evaluation Metrics: We use metrics of Dice (F-Measurement) and Jaccard (overlapping) to evaluate the segmentation performance of our model. An additional *average score*, which is the average of the Dice and Jaccard, is used as a balanced evaluation criterion. The metrics are defined as:

$$Dice(DC) = \frac{2 \times tp}{2 \times tp + fp + fn} \tag{3}$$

$$Jaccard(Jc) = \frac{tp}{tp + fp + fn} \tag{4}$$

$$average\ score = \frac{1}{2} \times (Dice + Jaccard) \tag{5}$$

where tp, fp, tn and fn refer to true positive, false positive, true negative, and false negative, respectively.

3) Implementation Details: Preprocessing procedures are conducted to all images before training, including centred cropping into the square shape, resizing into 512×512 by bilinear interpolation, transforming into grayscale and enhancement by contrast-limited adaptive histogram equalization [34]. Additional random flipping is utilized on images for data augmentation during training. We implemented our architecture with MobileNetV2 [35] as the encoder in U-shape backbone, within PyTorch repository. The MobileNetV2 is pre-trained on ImageNet [36] and all the layers of the networks are later fine-tuned. All the experiments are performed using the Adam [37] optimizer with a fixed learning rate of 1e-4 and a batch size of 16. For weakly-supervised initiation, models are trained for ten epochs in total. For fully-supervised domain adaptation and the conventional fully-supervised segmentation training, the validation set is used for early-stopping. All experiments are conducted on a dual-GPU system with an NVIDIA GeForce GTX 1080 Ti and an NVIDIA GeForce RTX 2080 Ti. The average time for weakly-supervised initiation is around 45 minutes. The average time for fully-supervised domain adaptation varies from 10 to 30 minutes and for fully-supervised segmentation training ranges from 20 to 50 minutes, depending on the number of samples are used.

B. Ablation Studies

In this subsection, we first investigate the effectiveness of our proposed CSC within different training schemes. Then, more experiments are conducted specifically to evaluate our HyProcess approach.

1) Effectiveness of CSC: To demonstrate the effectiveness of the CSC, we conduct ablation experiments thoroughly on weakly-supervised, fully-supervised and HyProcesss training schemes.

- **With Weakly-Supervision Only** Due to the inadequate supervision signal of image-level labels, conventional weakly-supervised methods can barely find the approximate location with a large potential region of the OD, and the gridding artifacts are obvious (Fig. 5d, 5e, 5f, and 5g). Even with the help of deep supervision technique, the slightly improved results still appear inaccurate (Fig. 5d and 5f). Comparatively, the enriched multi-level visual semantics produced by CSC allows our network to better capture the detailed characteristics of the OD. Our model achieves the best result with the correct location and matched region (Fig. 5c), indicating a strong capability to separate the OD from the background.

- **With Fully-Supervision Only** Improvement can be obtained by simply embedding our CSC into the conventional fully-supervised network. As observed in fully-supervised part of Table II, CSC improves the average score by 0.75(%), 1.74(%) and 0.59(%) on ONHSD, DRIONS-DB, and DRISHTI-GS datasets than the backbone U-Net, such improvements indicate that CSC effectively brings enriched multi-scale visual semantics to the model and fuses them in a reasonable way.

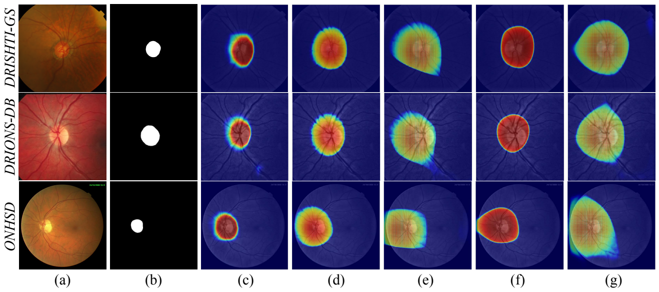


Fig. 5. Visual comparisons of score maps after weakly-supervised initiation with different models. (a) Source image; (b) Ground truth; (c) Results of ours (backbone + deep supervision + CSC); (d) Results of ours (backbone + deep supervision); (e) Results of ours (backbone) ; (f) Results of M-Net; (g) Results of BRU-Net.

TABLE II
ABLATION ANALYSIS ON AVERAGE SCORE (%) OF OUR APPROACH WITH EITHER FULLY-SUPERVISED OR HYBRID-SUPERVISED TRAINING PROTOCOLS. THE BEST RESULTS ARE SHOWN AS TEXT IN BOLD. **B** DENOTES BACKBONE, AND **CSC** DENOTES CROSS-SCALE CONNECTION.

Dataset	ONHSD		DRIONS-DB		DRISHTI-GS	
# of Samples	5	45	5	45	5	45
fully-supervised						
B	-	78.02	-	88.75	-	92.88
B + CSC	-	78.77	-	90.49	-	93.47
hybrid-supervised (HyProcess)						
B	74.12	79.24	77.75	90.03	85.48	93.75
B + CSC	**75.82**	**80.31**	**84.51**	**90.98**	**89.33**	**94.94**

- **With HyProcess** As seen in hybrid-supervised part of Table II, within both scenarios of given five and forty-five training samples, considerable improvements on average score can be found in models with the CSC. Specifically, the baseline is improved by 1.70(%), 6.76(%), 3.85(%) given five samples and by 1.07(%), 0.95(%), 1.19(%) given forty-five samples on three public datasets. Our CSC improves performance with fewer samples and yields the best segmentation results, suggesting that even with only a few pixel-level annotations, our CSC is still able to extract sufficient visual information.

Evidently, the CSC helps capture more spatial features and achieves better performance in all cases. Specifically, since the models under the weakly-supervision procedure are only trained with limited supervised signals in HyProcess, such feature-sharing mechanism is necessary to preserve sufficient spatial semantics for the weakly-initiation, as shown in Fig.5.

2) Effectiveness of HyProcess: As shown in Table II, the HyProcess brings improvement of 1.22(%), 1.28(%) and 0.87(%) to the U-Net backbone, and still brings improvement of 1.54(%), 0.49(%) and 1.47(%) to model with both backbone and the CSC, compared to the fully-supervised scheme. Furthermore, even with a weakly-supervised initiation on models with SPPH, no sign of performance degradation due to domain shift is observed. Meanwhile, their performance is enhanced, indicating that HyProcess prevents the worsening of the do-

TABLE III
DICE RESULTS OF HYNET COMPARED WITH THE STATE-OF-THE-ART UNDER HYPROCESS SCHEME. BEST RESULTS ARE SHOWN IN BOLD.

Dataset	ONHSD		DRIONS-DB		DRISHTI-GS	
# of Samples	5	45	5	45	5	45
Deep Disc [38]	71.30	80.37	85.08	92.59	87.68	93.30
M-Net [7]	76.78	81.56	85.14	93.13	89.97	94.94
CE-Net [9]	78.32	79.74	84.59	91.93	87.39	94.64
U-Net [15]	78.34	82.33	83.39	93.03	89.58	95.13
BRU-Net [8]	78.25	81.93	83.40	93.17	89.63	95.21
HyNet (ours)	**79.72**	**82.39**	**88.82**	**93.72**	**92.47**	**95.34**

TABLE IV
DICE RESULTS OF HYNET COMPARED WITH THE STATE-OF-THE-ART UNDER THE FULLY-SUPERVISED SCHEME. ALL MODELS ARE GIVEN 45 PIXEL-LEVEL ANNOTATED SAMPLES. BEST RESULT ARE SHOWN IN BOLD.

Model	ONHSD	DRIONS-DB	DRISHTI-GS
DeepDisc	81.11	92.93	94.13
M-Net	81.61	91.85	94.57
CE-Net	81.10	90.97	93.65
U-Net	81.36	92.04	95.04
BRU-Net	81.34	93.31	95.27
HyNet (ours)	**81.97**	**93.36**	**95.47**

main shift problem and greatly improves the segmentation performance of all models. The results of the model under the fully supervised scenario with only five samples are unstable due to severe overfitting and therefore not be given.

C. Comparisons to State-of-the-Arts

In this subsection, we compare our HyNet with five previous state-of-the-art methods, including U-Net [15], M-Net [7], DeepDisc [38], CE-Net [9], and BRU-Net [8]. We reimplement some of the above models with MobileNetV2 encoder. Results of all models are generated without any post-processing tools and evaluated with the same code.

1) Quantitative Result: Quantitative results of Dice are listed in Table III and Table IV for comparisons in three public datasets within hybrid- and fully-supervised schemes, respectively. Firstly, as shown in Table III, our HyNet surpasses all previous state-of-the-art methods under the HyProcess scheme, and achieves the best performance in both cases given five and forty-five training samples. As shown in Table IV, our HyNet also produces the best results within the conventional fully-supervised scheme. The promising performance within both hybrid- and fully-supervised schemes demonstrates the advancement of our architecture for OD segmentation.

2) Visual Comparisons: To further explain the advantages of our method, we show some qualitative results. As shown in Fig. 6, our approach yields promising segmentation results with only five pixel-level annotated samples, and is no longer plagued by the domain-shift problem. Compared to other state-of-the-art models trained with abundant samples, our results are more accurate and solid.

TABLE V
COMPARISONS OF PARAMETERS AMOUNT, INFERENCE TIME(MS) AND SEGMENTATION PERFORMANCE (DC), GIVEN FORTY-FIVE SAMPLES UNDER HYPROCESS SCHEME. **B** DENOTES BACKBONE, AND **CSC** DENOTES CROSS-SCALE CONNECTION.

Model	# Param	DRISHTI-GS		DRIONS-DB		ONHSD	
		Time↓	DC↑	Time↓	DC↑	Time↓	DC↑
U-Net [15]	32.2M	31.45	95.13	30.51	93.03	30.98	82.33
DeepLabV3+ [24]	5.80M	10.69	95.21	10.35	93.32	10.47	82.35
HyNet (B)	**5.01M**	**7.89**	94.01	**7.47**	92.98	**7.56**	82.26
HyNet (B + CSC)	5.30M	8.95	**95.34**	8.56	**93.72**	8.72	**82.39**

D. Further Analysis

1) Does the Number of Pixel-level Samples Matter?: We demonstrate the Jaccard results of U-Net and HyNet trained with different numbers of pixel-level annotated samples, under fully-supervised and HyProcess training schemes. As shown in Fig. 7, observations are twofold: **Firstly**, models trained under HyProcess outperform models trained under the fully-supervised scheme, especially when the models are given a limited samples (like five or ten), indicating the effectiveness of weakly-supervised initiation in a very small data regime. **Secondly**, the performance of all models improves as the number of samples increased. It is worth noting that the performance of the HyProcess models and the fully-supervised models becomes close when the number of samples increases to more than thirty, which indicates that pixel-level annotations contain more detailed visual information than image-level ones. Only given sufficient samples can the fully-supervised models be comparable to our HyProcess models.

Evidently, in all cases, the number of pixel-level annotated samples is critical, and it is particularly sensitive in conventional fully-supervised scenarios. Thanks to weakly-supervised initiation, our method achieves promising performance even with very limited pixel-level annotated samples, improves stably as more samples become available, and maintains the best performance compared to other methods.

2) Why are HyNet and HyProcess Important for Clinical Practice?: To assess the importance of HyNet and HyProcess for clinical OD segmentation, two observations are critical: **Firstly**, by comparing the results on all three datasets in Table III and Table IV, our HyNet, which is given only five samples under HyProcess scheme, performs comparably to the fully-supervised model, which is given forty-five pixel-level annotated samples. Even with very limited samples, our method still produces promising results and overcomes the domain shift problem. Evidently, our method provides a much more cost-efficient way to establish a clinical diagnosis system across different medical facilities than previous systems. **Secondly**, as shown in Fig. 7, our approach yields the best performance with a large number of samples, demonstrating that it consistently provides the best segmentation performance for healthcare institutes, regardless of whether they have a limited or large amount of data.

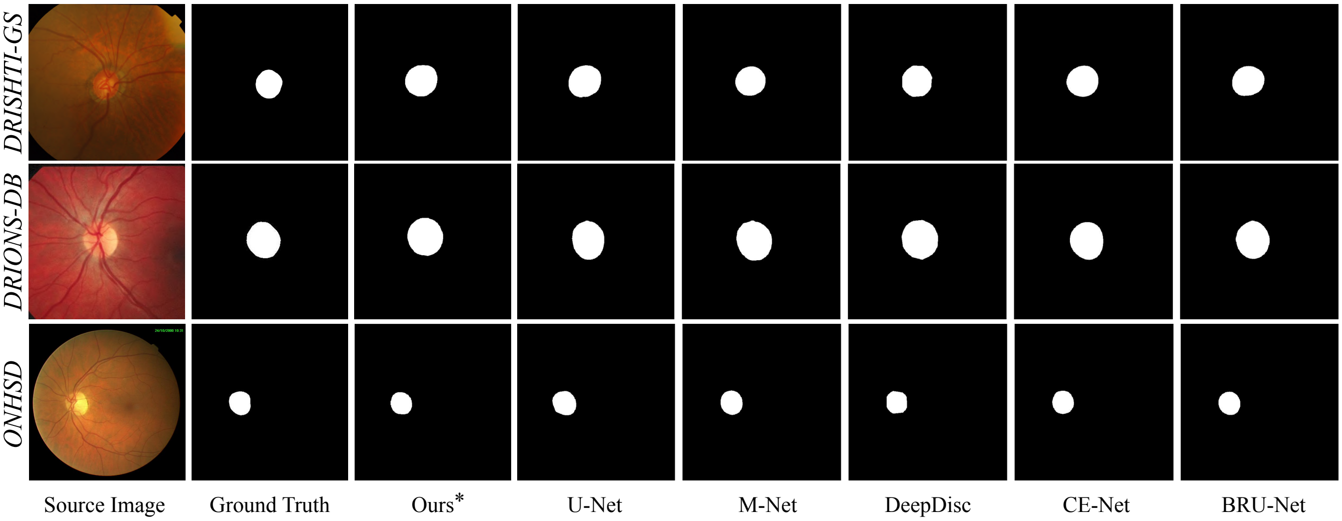


Fig. 6. Qualitative comparisons to state-of-the-art methods. ∗: Note that our model is trained with HyProcess scheme given only five pixel-level annotated samples from each dataset, while the others are trained with fully-supervised scheme given forty-five pixel-level samples.

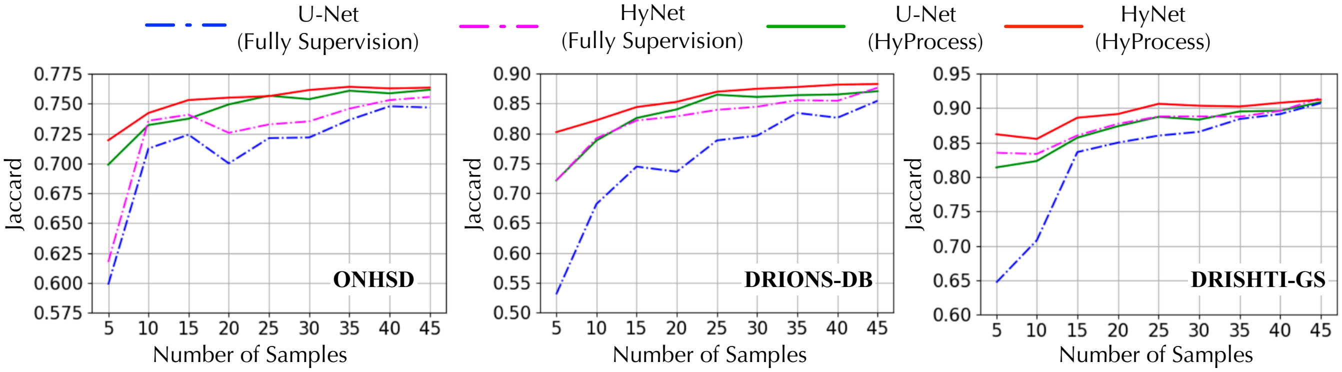


Fig. 7. Jaccard results of U-Net and HyNet given different number of samples, under fully-supervised and HyProcess schemes.

In summary, in clinical practice, our method helps to reduce the reliance on pixel-level annotations, overcome the domain shift problem, and obtain the best segmentation results at the same time.

3) Efficiency and Speed: We compare the computational efficiency of our model with the original U-Net and the latest DeepLabV3+ [24], as shown in Table V. Comparatively, even with fewer parameters, our HyNet obtains the best results of Dice than previous methods and maintains the fastest inference speed.

V. Conclusion

In this paper, we propose the HyNet with the HyProcess scheme as a practical tool for OD segmentation. Firstly, we reveal that our HyNet with CSC structure successfully captures information-rich features within the weakly-initiation procedure. Later, through utilizing the HyProcess scheme, our model achieves promising results with only a small number of pixel-level annotated samples and successfully overcomes the issue of domain shift. Extensive experiments show that our method surpasses the state-of-the-arts on three public OD segmentation datasets. Moreover, we analyze the inference speed of our method, and the results demonstrate it can serve as an accurate, robust, efficient and economical solution on OD segmentation in clinical practice.

References

[1] H. A. Quigley and A. T. Broman, "The number of people with glaucoma worldwide in 2010 and 2020," *British journal of ophthalmology*, vol. 90, no. 3, pp. 262–267, 2006.

[2] F. Yin, J. Liu, D. W. K. Wong, N. M. Tan, C. Cheung, M. Baskaran, T. Aung, and T. Y. Wong, "Automated segmentation of optic disc and optic cup in fundus images for glaucoma diagnosis," in *2012 25th IEEE international symposium on computer-based medical systems (CBMS)*. IEEE, 2012, pp. 1–6.

[3] A. M. Jose and A. A. Balakrishnan, "A novel method for glaucoma detection using optic disc and cup segmentation in digital retinal fundus images," in *2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015]*. IEEE, 2015, pp. 1–5.

[4] M. Abramoff, W. Alward, E. Greenlee, L. Shuba, C. Kim, J. Fingert, and Y. Kwon, "Automated segmentation of the optic nerve head from stereo color photographs using biologically plausible feature detectors," *Investigative Ophthalmology and Visual Sciences*, 2007.

[5] B. Al-Bander, B. Williams, W. Al-Nuaimy, M. Al-Taee, H. Pratt, and Y. Zheng, "Dense fully convolutional segmentation of the optic disc and cup in colour fundus for glaucoma diagnosis," *Symmetry*, vol. 10, no. 4, p. 87, 2018.

[6] J. Son, S. J. Park, and K.-H. Jung, “Towards accurate segmentation of retinal vessels and the optic disc in fundoscopic images with generative adversarial networks,” *Journal of digital imaging*, vol. 32, no. 3, pp. 499–512, 2019.

[7] H. Fu, J. Cheng, Y. Xu, D. W. K. Wong, L. Jiang, and X. Cao, “Joint optic disc and cup segmentation based on multi-label deep network and polar transformation,” *IEEE Transactions on Medical Imaging*, vol. PP, no. 99, pp. 1–1, 2018.

[8] S. Apostolopoulos, S. D. Zanet, C. Ciller, S. Wolf, and R. Sznitman, “Pathological oct retinal layer segmentation using branch residual u-shape networks,” 2017.

[9] Z. Gu, J. Cheng, H. Fu, K. Zhou, and J. Liu, “Ce-net: Context encoder network for 2d medical image segmentation,” *IEEE Transactions on Medical Imaging*, vol. PP, no. 99, pp. 1–1, 2019.

[10] S. Wang, L. Yu, X. Yang, C.-W. Fu, and P.-A. Heng, “Patch-based output space adversarial learning for joint optic disc and cup segmentation,” *arXiv preprint arXiv:1902.07519*, 2019.

[11] X. Feng, J. Yang, A. F. Laine, and E. D. Angelini, “Discriminative localization in cnns for weakly-supervised segmentation of pulmonary nodules,” in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2017, pp. 568–576.

[12] Y. Xu, J. Y. Zhu, I. C. Chang, M. Lai, and Z. Tu, “Weakly supervised histopathology cancer image segmentation and classification,” *Medical Image Analysis*, vol. 18, no. 3, pp. 591–604, 2014.

[13] Z. Jia, X. Huang, I. Eric, C. Chang, and Y. Xu, “Constrained deep weak supervision for histopathology image segmentation,” *IEEE transactions on medical imaging*, vol. 36, no. 11, pp. 2376–2388, 2017.

[14] B. Zou, Q. Liu, K. Yue, Z. Chen, J. Chen, and G. Zhao, “Saliency-based segmentation of optic disc in retinal images,” *Chinese Journal of Electronics*, vol. 28, no. 1, pp. 71–75, 2019.

[15] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[16] Q. Dou, C. Ouyang, C. Chen, H. Chen, and P.-A. Heng, “Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss,” *arXiv preprint arXiv:1804.10916*, 2018.

[17] M. Javanmardi and T. Tasdizen, “Domain adaptation for biomedical image segmentation using adversarial training,” in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 554–558.

[18] T. Karras, T. Aila, S. Laine, and J. Lehtinen, “Progressive growing of gans for improved quality, stability, and variation,” *arXiv preprint arXiv:1710.10196*, 2017.

[19] K. Roth, A. Lucchi, S. Nowozin, and T. Hofmann, “Stabilizing training of generative adversarial networks through regularization,” in *Advances in neural information processing systems*, 2017, pp. 2018–2028.

[20] M. Oquab, L. Bottou, I. Laptev, and J. Sivic, “Is object localization for free?-weakly-supervised learning with convolutional neural networks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 685–694.

[21] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” 2015.

[22] Z. Lu and D. Chen, “Weakly supervised and semi-supervised semantic segmentation for optic disc of fundus image,” *Symmetry*, vol. 12, no. 1, p. 145, 2020.

[23] R. Zhao, W. Liao, B. Zou, Z. Chen, and S. Li, “Weakly-supervised simultaneous evidence identification and segmentation for automated glaucoma diagnosis,” vol. 33, no. 01, pp. 809–816, 2019.

[24] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[25] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[26] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, “A simple pooling-based design for real-time salient object detection,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3917–3926.

[27] F. Yu, V. Koltun, and T. Funkhouser, “Dilated residual networks,” 2017.

[28] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, “Unet++: A nested u-net architecture for medical image segmentation,” in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Springer, 2018, pp. 3–11.

[29] K. C. L. Wong, M. Moradi, H. Tang, and T. Syedamahmood, “3d segmentation with exponential logarithmic loss for highly unbalanced object sizes,” pp. 612–619, 2018.

[30] D. Mishra, S. Chaudhury, M. Sarkar, and A. S. Soin, “Ultrasound image segmentation: A deeply supervised network with attention to boundaries,” *IEEE Transactions on Biomedical Engineering*, vol. 66, no. 6, pp. 1637–1648, 2019.

[31] E. J. Carmona, M. Rincón, J. García-Feijoó, and J. M. Martínez-De-La-Casa, “Identification of the optic nerve head with genetic algorithms,” *Artificial Intelligence in Medicine*, vol. 43, no. 3, pp. 243–259, 2008.

[32] J. Sivaswamy, S. R. Krishnadas, G. D. Joshi, M. Jain, and A. U. S. Tabish, “Drishti-gs: Retinal image dataset for optic nerve head(onh) segmentation,” in *IEEE International Symposium on Biomedical Imaging*, 2014.

[33] J. Lowell, A. Hunter, D. Steel, A. Basu, R. Ryder, E. Fletcher, and L. Kennedy, “Optic nerve head segmentation.” *IEEE Trans Med Imaging*, vol. 23, no. 2, pp. 256–264, 2004.

[34] J. Skilling and S. F. Gull, “Algorithms and applications,” *Lecture Notes in Computer Science*, vol. 14, no. 3, pp. 83–132, 2010.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “Mobilenetv2: Inverted residuals and linear bottlenecks,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[36] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, and M. Bernstein, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[37] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *Computer Science*, 2014.

[38] Z. Gu, P. Liu, K. Zhou, Y. Jiang, H. Mao, J. Cheng, and J. Liu, “Deepdisc: Optic disc segmentation based on atrous convolution and spatial pyramid pooling,” in *Computational Pathology and Ophthalmic Medical Image Analysis*. Springer, 2018, pp. 253–260.