HACDR-Net: Heterogeneous-Aware Convolutional Network for Diabetic Retinopathy Multi-Lesion Segmentation

QiHao Xu[1, 2], Xiaoling Luo[1, 2*], Chao Huang[3], Chengliang Liu[2], Jie Wen [2], Jialei Wang[2], Yong Xu[2]

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen, China
[2]Shenzhen Key Laboratory of Visual Object Detection and Recognition, Harbin Institute of Technology, Shenzhen, China
[3]School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University, Shenzhen 518107, China
xqh51199597@outlook.com, xiaolingluoo@outlook.com, huangch253@mail.sysu.edu.cn, liucl1996@163.com, jiewenpr@126.com, 380468802@qq.com, laterfall@hit.edu.cn

Abstract

Diabetic Retinopathy (DR), the leading cause of blindness in diabetic patients, is diagnosed by the condition of retinal multiple lesions. As a difficult task in medical image segmentation, DR multi-lesion segmentation faces the main concerns as follows. On the one hand, retinal lesions vary in location, shape, and size. On the other hand, because some lesions occupy only a very small part of the entire fundus image, the high proportion of background leads to difficulties in lesion segmentation. To solve the above problems, we propose a heterogeneous-aware convolutional network (HACDR-Net) that composes heterogeneous cross-convolution, heterogeneous modulated deformable convolution, and optional near-far-aware convolution. Our network introduces an adaptive aggregation module to summarize the heterogeneous feature maps and get diverse lesion areas in the heterogeneous receptive field along the channels and space. In addition, to solve the problem of the highly imbalanced proportion of focal areas, we design a new medical image segmentation loss function, Noise Adjusted Loss (NALoss). NALoss balances the predictive feature distribution of background and lesion by jointing Gaussian noise and hard example mining, thus enhancing awareness of lesions. We conduct the experiments on the public datasets IDRiD and DDR, and the experimental results show that the proposed method achieves better performance than other state-of-the-art methods. The code is open-sourced on github.com/xqh180110910537/HACDR-Net.

Introduction

Diabetic retinopathy (DR) is one of the most common microvascular complications of diabetes, which can cause a series of fundus lesions. Therefore, DR multi-lesion segmentation is crucial to diabetes diagnosis. Over the past few years, Convolutional Neural Networks (CNNs) and Transformer Networks (Liu et al. 2023a,b) have greatly promoted the development of DR multi-lesion segmentation (Wang et al. 2022; Cui et al. 2023; Xu et al. 2022; Ling et al. 2023). However, existing segmentation methods still face limitations that hinder their performance in DR multi-lesion segmentation. First, each type of lesion has a variable shape and size in the fundus image. Secondly, the area occupied by the lesions in the fundus picture is small. As a result, model training is more likely to be biased toward the background rather than the lesions. Existing methods improve segmentation accuracy in DR multi-lesion segmentation by obtaining global position relationships through large receptive fields, but small lesion features are negatively affected by large irrelevant areas.

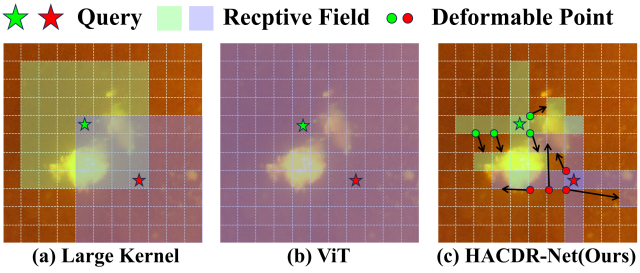

Figure 1: Comparison of our model and other attention patterns. (a) is a convolutional attention mechanism with very large kernels. (b) is ViT's (Dosovitskiy et al. 2020) global attention mechanism. (c) is our heterogeneous convolutional attention mechanism.

In this paper, we propose a heterogeneous-aware convolutional network for DR multi-lesion segmentation (HACDR-Net). Heterogeneous convolution aims to aggregate the convolution features of different structures to obtain heterogeneous receptive fields. Compared with previous DR multi-lesion segmentation methods, the heterogeneous receptive field extracts the heterogeneous features of lesions, thereby having a good segmentation effect on lesions of various sizes and shapes. Furthermore, the heterogeneous convolutional structure has space-adaptive capabilities to reduce perturbance in large irrelevant regions.

Aggregating heterogeneous convolution information is difficult, because features may conflict under heterogeneous receptive fields. To this end, inspired by Visual Attention Network (Guo et al. 2023) (VAN), we design a heterogeneous-aware attention aggregation (HAAA) module to summarize the heterogeneous feature maps. Different from VAN, we aggregate the features of heterogeneous convolution instead of a single large kernel convolution. In

*Corresponding Author: Xiaoling Luo
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Fig. 1, the receptive field of (a) and (b) is too large, causing the lesion to be easily perturbed by the background area; the heterogeneous receptive field of (c) is easier to get diverse lesion areas.

In addition, another difficulty is that the proportion of the focal area is imbalanced, which will cause the features of some lesions to be ignored during training. This is not allowed. We found that adjusting the predicted pixel values can change the distribution of predictive features. To this end, we propose a novel loss function, Noise Adjusted Loss (NALoss). NALoss balances the predictive feature distribution by adjusting the pixel prediction. Gaussian noise is added in the predicted pixels to perturb the background's feature and enhance the lesion's feature. It is worth noting that noise addition is not involved in the testing phase. Experiments prove that NALoss can strengthen the feature learning, and improves the segmentation performance. To summarize, our contributions are as follows:

- We propose a novel heterogeneous-aware convolutional network (HACDR-Net) for DR multi-lesion segmentation. The network has heterogeneous receptive fields and spatial adaptability, which solves the segmentation problem caused by the different shapes and sizes of lesions.
- We propose a new loss function Noise Adjusted Loss (NALoss) specially designed for fitting highly imbalanced segmentation. It balances the distributions of predictive features by jointing Gaussian noise and hard example mining.
- HACDR-Net undergoes thorough testing on two datasets and consistently achieves state-of-the-art results. Various metrics have been significantly improved on the DDR and IDRiD datasets.

Related Work

Approaches of DR Multi-Lesion Segmentation

In medical image segmentation, U-net and its family, such as ResUnet (Diakogiannis et al. 2020), DenseUnet (Li et al. 2018), and Unet++ (Zhou et al. 2019b), were first widely used. L-seg (Guo et al. 2019) first proposes an end-to-end unified framework for multi-lesion segmentation of fundus images. However, these performed poorly in DR multi-lesion segmentation. Because the receptive field of traditional convolution is too small, it is not enough to grasp the global relationship.

In recent years, segmentation models based on Transformer and CNN-Transformer, such as Transunet (Chen et al. 2021) and Swinunet (Cao et al. 2022), have begun to be applied to DR multi-lesion segmentation. Among them, RTnet (Huang et al. 2022) proposed a relation transformer network for diabetic retinopathy multi-lesion segmentation. The segmentation network with Swin Transformer (Liu et al. 2022) and Twins-SVT (Chu et al. 2021) as the backbone also achieved good results. PMCNet (He et al. 2022) improves the accuracy through the combination of CNN and Transformer. M2MRF (Liu et al. 2023c) is also a state-of-the-art network in this task. These networks mainly improve the segmentation effect by expanding the receptive field. But they neglected the characteristics of lesions and appeared helpless when facing sundry lesions.

Loss Function for Imbalanced Segmentation

Various segmentation loss functions for solving imbalanced medical image data problems have been widely used. There are two types of loss functions. The first type aims to balance the importance of samples. Examples include Weighted Cross-Entropy (Ronneberger, Fischer, and Brox 2015), Diceloss (Li et al. 2020), Focalloss (Lin et al. 2017). The second type aims to balance the number of samples, and one method for achieving this is online hard example mining (OHEM) (Wang et al. 2023).

Method

Overview of Our Work

The overall architecture of our proposed Heterogeneous-Aware Convolutional Network (HACDR-Net) is illustrated in Fig. 2, including HACDR-Net and Noise Adjusted Loss (NALoss). The encoder comprises four stages, each with downsampling rates $R_i = [4, 8, 16, 32]$. Each stage extracts heterogeneous features through repeatedly Heterogeneous Convolutional Attention (HCA) Blocks and downsamples with modulated deformable convolution (MDConv). The number of HCA Block iteration in the four stages respectively are 3, 3, 5, and 2. The core of HCA Block is the heterogeneous-aware attention aggregation (HAAA) module. For the deformable feed-forward network (DFFN) module in HCA block, we try to replace depth-wise convolution (DWConv) with MDConv. The decoder adopts a U-shaped structure like U-net (Ronneberger, Fischer, and Brox 2015). All of the structures in our HACDR-Net are residuals. In addition, we propose a new medical segmentation loss function NALoss for lesion-sample training, as shown in Fig. 4. It adjusts the feature distribution of training predictions by jointing Gaussian noise and hard example mining.

Encoder with HCA Block

As shown in Fig. 2, our encoder adopts a Transformer-like architecture, including deformable convolutional downsampling and HCA Block. However, different from self-attention and multi-head attention, we propose a novel heterogeneous convolutional attention to meet the requirements of lesion segmentation. In HCA Block, HAAA obtains heterogeneous feature maps through multi-branch heterogeneous convolution and then aggregates these features through an attention method. This mechanism of heterogeneous convolution can obtain diverse lesion areas in the heterogeneous receptive field along the channels and space. HCA Block widely uses MDConv. Moreover, a 3×3 MDConv is also applied in the downsampling. MDConv can grasp the details of various lesions and dynamically adapt to heterogeneous features, compared with traditional convolution. The MDConv is defined as follows:

$$\widetilde{F}(x, y) = \sum_{k=1}^{K} w_k m_k * F(x + \Delta x_k, y + \Delta y_k), \quad (1)$$

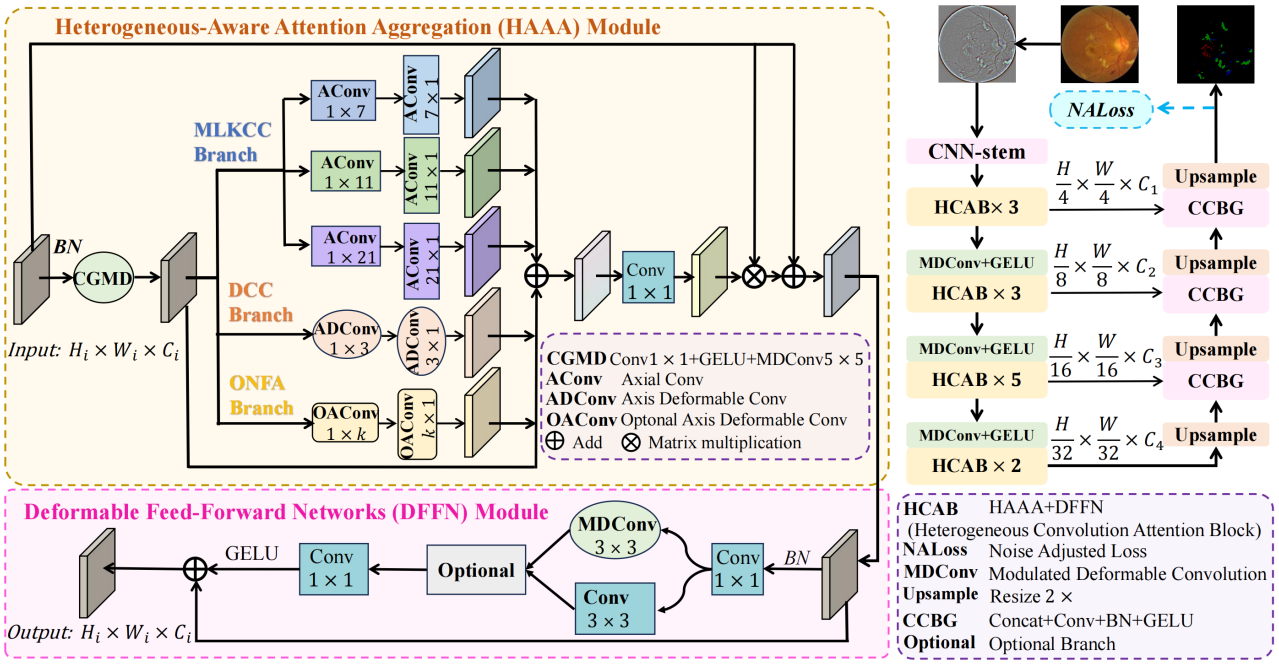

Figure 2: An overview of HACDR-Net.

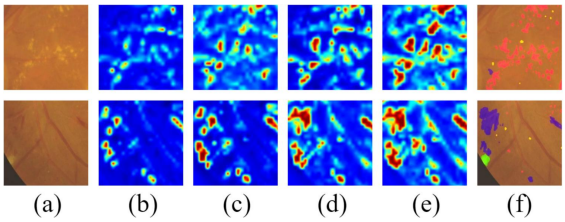

Figure 3: Visualization of the encoder's features by Grad-CAM (Selvaraju et al. 2017). (a) and (f) describes the original image and ground truth, respectively. (b) to (e) feature maps represent different branch structures and their combinations. (b) represents DCC, (c) represents MLKCC+ONFA, (d) represents DCC+MLKCC+ONFA, and (e) represents DCC+MLKCC+ONFA through the attention aggregation module.

where $F(x, y)$ represents the input features, and $\widetilde{F}(x, y)$ represents the deformed and enhanced features, where K represents the total number of sampling points, and k enumerates the sampling points. w_k represents the learnable weight of the k-th sampling point, and m_k represents the scalar modulation of the k-th sampling point, normalized by the sigmoid function. $(x + \Delta x_k, y + \Delta y_k)$ represents the offset coordinates of the sampling point. In this way, our HACDR-Net enhances awareness of lesions.

HAAA Module. To solve the problem of segmentation caused by various shapes and sizes of lesions, we use multi-branch heterogeneous convolution to obtain heterogeneous features, as shown in Fig. 2. Before extracting heterogeneous features, HAAA module uses 5×5 MDConv for feature dynamic adaptation.

We design three heterogeneous convolution branches, including a multi-scale large-kernel cross-convolution (MLKCC) branch, a deformable cross-convolution (DCC) branch, and an optional near-far-aware (ONFA) branch. As illustrated in Fig. 2, k×k cross-convolution is the convolution of features sequentially through 1×k and k×1 axis-convolutions. We use cross-convolution extensively here to reduce the collision of large irrelevant areas while obtaining and enlarging heterogeneous receptive fields. MLKCC branch obtains multi-scale cross receptive fields to capture the long-range relationship of lesions. Obtaining only a single-shaped cross receptive field cannot adapt to the characteristics of various shapes and sizes of lesions. DCC branch uses deformable cross-convolution to obtain dynamic local receptive fields. At the same time, DCC branch and 5×5 MDConv constitute a deformable convolution residual structure, which has a dynamic adaptive ability to get lesion areas of different shapes. ONFA branch is composed of k×k deformable cross-convolution, which can enhance network's adaptability. We residually sum these features to form a heterogeneous feature map.

Ultimately, an aggregation of heterogeneous feature channels is achieved through feature attention aggregation using 1×1 convolutions, culminating in an attention operation via input and output matrix multiplication. HAAA can be denoted as:

$$\begin{aligned} HAAA(X) = Conv_{1\times1}(\sum_{i=0}^{2} LKCC_{k_i\times k_i}(X') \\ + DCC_{3\times3}(X') + ONFA_{n\times n}(X') \\ + X') \otimes X + X, \end{aligned} \tag{2}$$

$$X' = MDConv_{5\times5}(X), \tag{3}$$
$$LKCC_{k_i\times k_i}(X) = DWConv_{1\times k_i}(DWConv_{k_i\times1}(X)), \tag{4}$$
$$DCC_{3\times3}(X) = MDConv_{1\times3}(MDConv_{3\times1}(X)), \tag{5}$$
$$ONFA_{n\times n}(X) = MDConv_{1\times n}(MDConv_{n\times1}(X)). \tag{6}$$

Among them, DWConv means depth-wise depth separable convolution, MDConv represents modulated deformable convolution, LKCC represents large-kernel cross-convolution, DCC indicates deformable cross-convolution, ONFA means optional near-far-aware convolution and their subscripts indicate the size of the convolution. The operation '$\otimes$' means element-wise matrix multiplication. We set the scale of axis convolution k_i to 7, 11, and 21.

Through multi-branch heterogeneous convolution and attention aggregation, HAAA possesses heterogeneous receptive fields and dynamic adaptive perception capability. HAAA reduces the feature collision of multi-branch and enhances awareness of various lesions. The three branches of MLKCC, DCC, ONFA, and the attention aggregation module are indispensable. As depicted in Fig. 3, the effect of heterogeneous convolution aggregation is significantly superior to that of other single branches.

DFFN Module. This module mainly enhances the local features of HAAA. As shown in Fig. 2, it adopts a residual convolution structure and can choose the form of MDConv and DWConv. It turns out that the two structures behave differently on different datasets.

Decoder

According to the requirements of DR multi-lesion segmentation, we choose the U-net structure network. We combine the multi-scale features of the encoder to form a feature map by channel fusion and upsampling. Finally, it is restored to a mask map. The decoder based on U-net (Ronneberger, Fischer, and Brox 2015) shows the best effect in DR multi-lesion segmentation.

Loss Function

To solve the problem of the highly imbalanced proportion of focal areas, we propose a loss function NALoss. As shown in Fig. 4, by jointing Gaussian noise and hard example mining, NALoss balances the predictive feature distribution of background and lesion to improve the feature representations.

First, during training, we add weighted Gaussian noise to the predicted pixels. By adjusting the distribution of predicted values for each pixel, NALoss can balance the distribution of predictive features.

Each predicted pixel is a vector of c categories and each mask pixel is the one-hot vector of c categories. We denote p_i as the predicted pixel, g_i as the mask pixel, and p_i^k, g_i^k as the k-th category value of the pixel. z_i^k denotes a special vector of the form p_i^k. w^k denotes the loss weight of category k. Adjusted Parameter can be denoted as α, $\boldsymbol{\alpha} = [\alpha^1, \alpha^2, \ldots, \alpha^k]$, where $\boldsymbol{\alpha^k}$ denotes the most critical noise weight of the k-th category. $\mathcal{N}(0, \sigma^2)$ is Gaussian noise with mean 0 and variance σ^2. Our loss formula for the first step can be denoted as $\mathcal{L}_{\mathcal{F}}$:

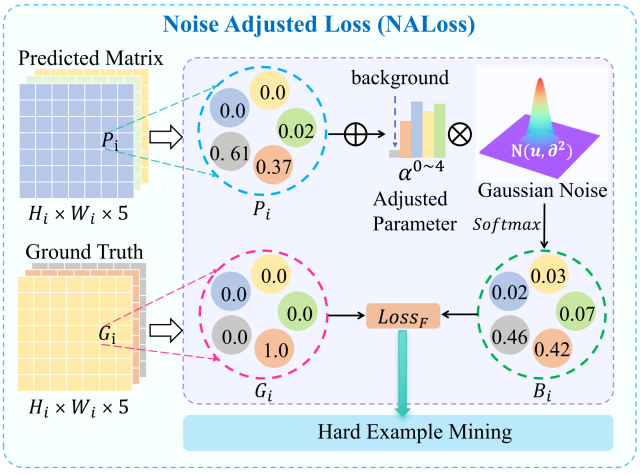

Figure 4: An overview of NALoss. The operation '$\oplus$' means add, and '$\otimes$' denotes multiply. Pick pixel P_i as an illustration to depict the dual stages of NALoss: adding Gaussian noise to balance the predictive distribution and hard example mining.

$$Softmax(z_i^k) = \frac{e^{z_i^k}}{\sum_{u=1}^{c} e^{z_i^u}}, \tag{7}$$

$$\xi = \log(Softmax(p_i^k + \boldsymbol{\alpha^k}\mathcal{N}(0, \sigma^2))), \tag{8}$$

$$\mathcal{L}_{\mathcal{F}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C} w^k g_i^k \xi. \tag{9}$$

Determining suitable Adjusted Parameter $\boldsymbol{\alpha}$ is the key to NALoss. To enhance the robustness of prediction for different pixels (background and lesion pixels), we collect the total number of pixels of different categories in the entire training set, and design the Adjusted Parameter $\boldsymbol{\alpha}$, as shown in the formula.

$$\boldsymbol{\alpha^k} = \frac{\log\frac{\sum_{j=1}^{C} s_j}{s_k}}{\sum_{i=1}^{C}\log\frac{\sum_{j=1}^{C} s_j}{s_i}}, \tag{10}$$

where s_k represents the total number of k-th category pixels of all images in the training set, $\sum_{j=1}^{C} s_j$ represents the total number of pixels in all images from the training set. We can see that Adjusted Parameter $\boldsymbol{\alpha}$ is inversely proportional to the number of categories, as shown in Fig. 4. That is to say, we will increase the prediction probability of lesions to achieve the effect of perturbing the background and increasing the proportion of lesion-predictive feature distribution.

Gaussian noise can cause some random variation in predicted ranges. With appropriate Adjusted Parameter $\boldsymbol{\alpha}$, fluctuating features can increase the difficulty of background

Methods	IoU					Dice					AUPR				
	mIoU	EX	SE	HE	MA	mDice	EX	SE	HE	MA	mAUPR	EX	SE	HE	MA
HED (Xie and Tu 2015)	46.66	64.74	47.43	50.38	24.07	62.18	78.60	64.33	67.00	38.81	63.94	80.81	66.41	68.09	40.45
PSPNet (Zhao et al. 2017)	41.70	57.78	43.71	45.81	19.50	57.38	73.24	60.81	62.83	32.63	58.73	75.21	63.36	63.65	32.71
DenseUNet (Li et al. 2018)	46.71	66.51	45.57	45.86	30.55	62.56	79.89	62.61	62.89	44.83	65.06	81.01	66.10	67.10	46.01
Deeplabv3+ (Chen et al. 2018)	45.21	66.10	44.90	44.39	25.45	60.90	79.60	61.96	61.48	40.57	63.19	81.93	64.66	63.04	43.14
L-seg (Guo et al. 2019)	-	-	-	-	-	-	-	-	-	-	65.15	79.45	63.74	**71.13**	46.27
DNL (Yin et al. 2020)	42.28	57.67	44.80	47.03	19.61	57.94	73.15	61.87	63.96	32.78	59.09	75.12	64.04	64.73	32.48
HRNetV2 (Wang et al. 2020)	47.52	66.57	45.56	50.99	26.98	63.14	79.93	62.58	67.53	42.49	64.93	82.09	65.50	68.38	43.76
Twins-SVT-B (Chu et al. 2021)	47.07	64.68	44.91	**51.76**	26.92	62.79	78.56	61.98	**68.19**	42.42	63.84	80.09	63.12	68.86	43.27
TransUnet (Chen et al. 2021)	46.49	**67.76**	47.33	46.46	24.42	62.74	79.89	62.47	64.02	44.57	63.23	80.01	66.91	62.85	43.10
Swin-Unet (Cao et al. 2022)	47.76	66.26	48.36	47.54	28.86	63.53	79.71	65.19	64.43	44.79	64.48	81.34	66.57	64.91	45.10
Swin-Tv2 (Liu et al. 2022)	48.09	67.22	49.33	45.26	30.55	62.99	**80.12**	65.71	62.25	43.90	64.86	83.11	68.20	65.12	43.02
PMCNet (He et al. 2022)	43.12	-	-	-	-	56.02	-	-	-	-	68.08	**87.24**	71.11	67.05	46.94
M2MRF (Liu et al. 2023c)	48.56	66.07	48.58	48.16	**31.42**	64.45	79.57	65.39	65.01	47.81	66.00	81.98	67.41	66.68	**47.91**
HACDR-Net (Ours)	**49.12**	64.61	**56.21**	47.35	28.31	**64.71**	78.50	**71.96**	64.27	44.12	**68.79**	86.90	**76.73**	68.50	43.02

Table 1: Comparison of our proposed HACDR-Net with the state-of-the-art methods on the IDRiD dataset. The best results are highlighted in bold and the second best results are underlined. (Unit: %)

prediction, which also can strengthen the robustness of training.

Next, we propose a hard example mining of noise-adding pixels. The training focuses on the pixels where the noise-adding predictions are seriously wrong. The specific loss function NALoss is as follows:

$$\mathcal{L}_{NA} = \mathcal{L}_{\mathcal{F}}(\mathcal{L}_{\mathcal{F}} < \theta), \tag{11}$$

where θ represents the threshold. By hard example mining, NALoss significantly improves the segmentation performance of lesions.

As for the reason that we use hard example mining on noise-adding pixels, on the one hand, we found that the minor prediction loss caused by noise can be ignored. In this way, HACDR-Net can both focus on those lesion errors with low presence and reduce the negative impact of perturbations on the background. On the other hand, if we simply use hard example mining, the training will still be dominated by background pixels, which cannot solve the problems of training process. Attentively, we divide the training into two stages. In the initial stage, we use the Cross-Entropy loss function, and then employ NALoss for training.

Experiments

Datasets and Evaluation Metrics

Dataset. Two publicly available DR multi-lesion segmentation datasets are adopted, $i.e.$, the Indian Diabetic Retinopathy Image Dataset (IDRiD), A General-purpose High-quality Dataset for Diabetic Retinopathy Classification, Lesion Segmentation and Lesion Detection (DDR). These datasets consist of images with a background category and four kinds of lesion categories. The four types of lesions include hard exudates (EX), soft exudates (SE), microangiomas (MA), and hemorrhages (HE).

DDR: The DDR (Li et al. 2019) dataset contains 757 images of fundus lesions with pixel-level annotations, including 383 images for training, 149 images for validation, and 225 images for testing. The resolution of the images in this dataset ranges from 1088×1920 to 3456×5184 pixels.

IDRiD: The IDRiD (Porwal et al. 2018) dataset only contains 81 images of fundus lesions with pixel-level annotations, including 54 images for training and 27 images for testing. The resolution of the images in this dataset is 2848×4288 pixels.

Evaluation Metrics. We follow the protocol suggested by DDR (Li et al. 2019) and IDRiD (Porwal et al. 2018) and report standard metrics including Intersection over Union (IoU) (Rezatofighi et al. 2019), mean Intersection over Union (mIoU) (Rezatofighi et al. 2019), Dice coefficient (Milletari, Navab, and Ahmadi 2016), mean Dice coefficient (mDice) (Milletari, Navab, and Ahmadi 2016), the area under precision-recall curve (AUPR) (Boyd, Eng, and Page 2013) and mean area under precision-recall curve (mAUPR) (Boyd, Eng, and Page 2013). As multi-class segmentation tasks, mDice, mAUPR, and mIoU are core metrics for evaluating performance.

Implementation Details

Our implementation is based on mmsegmentation (Contributors 2020) libraries. All models are trained on a node with 2 RTX 3090 GPUs. Following M2MRF (Liu et al. 2023c), images in IDRiD are resized to 1440×960 pixels, and we resize the images of DDR to 1280×1280. To enhance the robustness of the model, we use three data augmentation techniques: multiple scaling (0.5∼2.0), rotation (90°, 180°, and 270°), and flipping (horizontal and vertical). Before training, we preprocess the images by contrast, brightness adjustment, and image fusion as used in (Zhou et al. 2019a). It can mitigate variation due to lighting conditions and resolution. The batch size is set to 1∼4 according to different resolutions for these two datasets. AdamW (Loshchilov and Hutter 2017) is applied to train our models. We set the initial learning rate as 0.00006 and employ the poly-learning rate decay policy.

Comparison with the State-of-the-Arts

Quantitative Comparison. We compare HACDR-Net with other state-of-the-art methods on the DDR and IDRiD

Methods	IoU					Dice					AUPR				
	mIoU	EX	SE	HE	MA	mDice	EX	SE	HE	MA	mAUPR	EX	SE	HE	MA
HED (Xie and Tu 2015)	27.17	39.50	27.09	29.46	12.63	41.79	56.63	42.61	45.50	22.43	42.97	61.40	43.19	46.68	20.61
PSPNet (Zhao et al. 2017)	24.31	37.31	24.51	26.64	8.75	37.97	54.35	39.37	42.08	16.09	39.23	57.04	42.71	42.32	14.85
DenseUNet (Li et al. 2018)	31.58	41.25	37.58	32.73	14.76	47.02	58.41	54.63	49.32	25.73	48.29	62.00	55.01	51.11	25.05
Deeplabv3+ (Chen et al. 2018)	26.47	41.44	23.44	26.46	14.55	40.95	58.59	37.97	41.83	25.40	42.34	62.32	40.79	41.83	24.39
L-seg (Guo et al. 2019)	-	-	-	-	-	-	-	-	-	-	32.08	55.46	35.86	26.48	10.52
DNL (Yin et al. 2020)	24.33	36.39	27.15	25.33	8.46	38.02	53.36	42.71	40.40	15.60	40.14	56.05	47.81	42.01	14.71
HRNetV2 (Wang et al. 2020)	28.84	41.82	29.01	28.94	15.60	43.95	58.98	44.96	44.86	26.99	45.21	61.55	45.68	46.91	26.70
Twins-SVT-B (Chu et al. 2021)	29.28	39.70	29.08	36.24	12.07	44.15	56.83	45.04	53.19	21.54	46.11	59.71	49.96	52.72	21.54
TransUnet (Chen et al. 2021)	27.78	39.76	37.43	22.46	11.47	42.15	56.89	54.47	36.68	20.57	44.03	60.57	55.46	41.49	18.58
Swin-Unet (Cao et al. 2022)	30.07	42.64	33.82	30.62	13.19	45.10	59.79	50.53	46.77	23.31	46.72	62.71	54.39	46.12	23.67
Swin-Tv2 (Liu et al. 2022)	32.10	44.07	36.12	32.77	15.44	47.86	61.18	54.14	49.36	26.75	48.59	63.05	55.01	50.11	26.20
PMCNet (He et al. 2022)	-	-	-	-	-	-	-	-	-	-	36.44	54.30	31.64	31.64	19.94
M2MRF (Liu et al. 2023c)	30.41	43.06	30.56	32.08	**15.95**	45.77	60.20	46.81	48.58	**27.51**	49.42	63.88	55.47	50.01	**28.33**
HACDR-Net (Ours)	**33.70**	**44.13**	**38.54**	**36.95**	15.17	**49.30**	**61.24**	**55.64**	**53.96**	26.34	**50.36**	**65.15**	**56.75**	**55.02**	24.50

Table 2: Comparison of our proposed HACDR-Net with the state-of-the-arts methods on the DDR dataset. The best results are highlighted in bold and the second best results are underlined. (Unit: %)

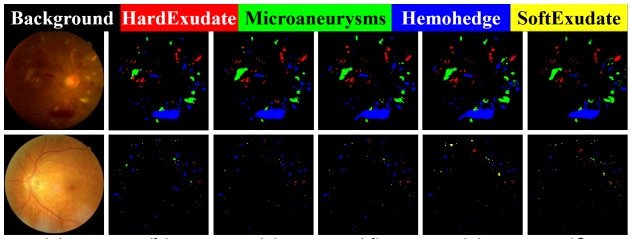

Figure 5: Visual Comparison of 4 methods on the DDR dataset. The colored boxes show the main lesions. (a) Fundus Image, (b) DenseUnet, (c) SwinV2, (d) M2MRF, (e) HACDR-Net (Ours), (f) Ground Truth.

dataset. These compared methods are mainly divided into three categories, including Convolutional networks, Transformer-based networks, and previous DR multi-lesion segmentation networks. Convolutional networks include HED (Xie and Tu 2015), PSPNet (Zhao et al. 2017), DenseUNet (Li et al. 2018), Deeplabv3+ (Chen et al. 2018), DNL (Yin et al. 2020), HRNetV2 (Wang et al. 2020). Transformer-Based networks include Swin-T-base (Liu et al. 2021), Twins-SVT-B (Chu et al. 2021), TransUnet (Chen et al. 2021), Swin-Unet (Cao et al. 2022), Swin-transformer V2 (Swin-Tv2) (Liu et al. 2022). Previous DR multi-lesion segmentation networks include L-seg (Guo et al. 2019), PMCNet (He et al. 2022), M2MRF (Liu et al. 2023c).

Table 1 lists the performance of different comparison methods on the IDRiD dataset. Likewise, compared with the previous best method M2MRF, our metrics mAUPR, mDice, and mIoU improve respectively by 2.79%, 0.26%, and 0.56%. Table 2 lists the performance of different comparison methods on the DDR dataset. Our HACDR-Net shows the best performance with all methods. Compared with the previous best method M2MRF, our metrics mAUPR, mDice, and mIoU improve respectively by 0.94%, 3.53%, and 3.29%. In the two datasets, not only the mean metrics but also the category metrics have improved. All in all, these

Methods	mDice	mIoU	mAUPR
$\mathcal{M}_{\mathcal{D}}$	46.55	31.04	46.50
$\mathcal{B}_{\alpha}$	47.35	32.01	48.01
$\mathcal{B}_{\theta}$	46.70	31.32	46.89
$\mathcal{B}_{\alpha} + \mathcal{M}_{\mathcal{D}}$	48.02	32.75	48.50
$\mathcal{B}_{\theta} + \mathcal{M}_{\mathcal{D}}$	48.45	33.04	49.21
$\mathcal{B}_{\alpha} + \mathcal{B}_{\theta} + \mathcal{B}_{\kappa} + \mathcal{M}_{\mathcal{D}}$	**49.30**	**33.70**	**50.36**
$\mathcal{B}_{\alpha} + \mathcal{B}_{\theta} + \mathcal{M}_{\mathcal{D}}$	49.04	33.26	49.92
w/o HAAA	45.31	30.43	45.70
w/o DFFN	49.09	33.42	49.70

Table 3: Ablation study in HACDR-Net on the DDR dataset. HACDR-Net overall is represented as '$\mathcal{B}_{\alpha}+\mathcal{B}_{\theta}+\mathcal{B}_{\kappa}+\mathcal{M}_{\mathcal{D}}$', where '$\mathcal{B}_{\alpha}$', '$\mathcal{B}_{\theta}$', '$\mathcal{B}_{\kappa}$', and '$\mathcal{M}_{\mathcal{D}}$' indicate MLKCC branch, DCC branch, ONFA branch, and the Modulated Deformable Convolution respectively. 'w/o HAAA' and 'w/o DFFN' denote that HAAA and DFFN are removed from the overall model, respectively. The best results are highlighted in bold. (Unit: %)

quantitative results on two datasets substantiate the fine robustness of our HACDR-Net.

Visual Comparison. Fig. 5 shows the qualitative results of different methods on the DDR dataset, including DenseUnet, M2MRF, Swin-transformer v2 (Swin-Tv2), and our HACDR-Net. It demonstrates that our HACDR-Net can precisely segment lesions of different shapes and sizes compared with other methods.

Ablation Study

Through experiments, we analyze the contributions of each component in our model, detailed in Table 3. We assess the model performance by removing HAAA, DFFN, and NALoss components. Additionally, we employ T-SNE (Van der Maaten and Hinton 2008) for feature visualization and conduct a hyperparameter analysis. Ablation experiments were conducted on two public datasets, with a focus on the more representative DDR dataset due to its diverse

Structure	mDice	mIoU	mAUPR
ψ	48.55	32.71	49.10
τ	48.02	32.01	48.01
υ	47.72	31.32	48.05
$\psi + \tau$	49.07	32.52	49.44
$\psi + \upsilon$	48.60	32.79	49.50
$\tau + \upsilon$	49.10	33.04	50.03
$\tau + \upsilon + \psi$	**49.30**	**33.70**	**50.36**

Table 4: Ablation study in HAAA on the DDR dataset. ψ is represented as 1×1 convolution, τ is represented as 5×5 modulated deformable convolution, and υ is represented as attention. The best results are highlighted in bold. (Unit: %)

Loss Function	mDice	mIoU	mAUPR
ζ_{CE}	47.04	31.67	47.32
ζ_{WCE}	47.88	32.34	48.51
ζ_{Dice}	46.50	30.81	46.51
$\zeta_{CE} + \zeta_{Dice}$	46.95	31.34	46.75
ζ_{OHEM}	47.12	31.71	47.95
$\zeta_{NA} + \zeta_{WCE}$	49.08	33.41	50.01
ζ_{NA}	**49.30**	**33.70**	**50.36**

Table 5: Ablation study of loss function on the DDR dataset. The top outcomes are shown in bold. (Unit: %)

image resolutions and qualities.

Effectiveness of Heterogeneous Branches. Table 3 shows that the heterogeneous features formed by HACDR-Net have greatly improved various metrics compared with a single branch. Among them, the combination of branches represented by '$\mathcal{B}_\alpha$', '$\mathcal{B}_\theta$', '$\mathcal{B}_\kappa$', compared with a single branch, our metrics mAUPR, mDice, and mIoU improve respectively by 2.35%$\sim$3.86%, 1.95%$\sim$2.75%, and 1.69%$\sim$2.66%. '$\mathcal{B}_\kappa$' ($k = 17$) has a significant biased effect on lesions of different scales. The details of '$\mathcal{B}_\kappa$' will be introduced in Hyper-Parameters. Overall, our results shed new light on the importance of heterogeneous-aware convolution for improving DR multi-lesion segmentation performance.

Effectiveness of HAAA and DFFN Module. HAAA collects and combines heterogeneous features in HACDR-Net. Table 3 demonstrates its significance, as mAUPR, mDice, and mIoU drop by 4.66%, 3.99%, and 3.27% without HAAA. Table 4 verifies the essentiality of HAAA's key components: 1×1 convolution for aggregation, 5×5 deformable convolution for adaptive features, and attention mechanism. DFFN is crucial for aggregating heterogeneous features and without it, mAUPR, mDice, and mIoU decrease by 0.66%, 0.21%, and 0.28%, respectively.

Effectiveness of NALoss. Table 5 illustrates that our NALoss outperforms the DR multi-lesion segmentation dataset with imbalanced data. Abbreviations used include ζ_{CE} for Cross-Entropy loss, ζ_{WCE} for Weighted Cross-Entropy loss (Ronneberger, Fischer, and Brox 2015), ζ_{Dice} for Dice loss (Li et al. 2020), ζ_{OHEM} for Ohem loss (Shrivastava, Gupta, and Girshick 2016), and ζ_{NA} for our

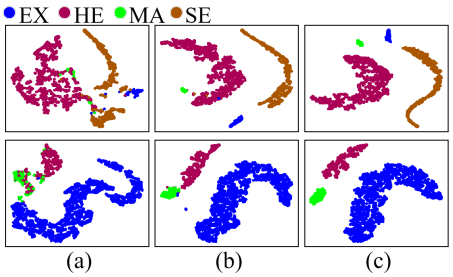

Figure 6: The deep features are visualized using T-SNE in (a) DenseU-net, (b) HACDR-Net+WCEloss, and (c) HACDR-Net+NALoss.

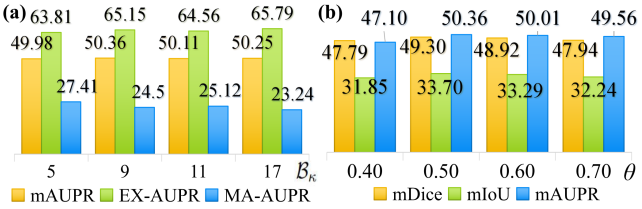

Figure 7: Evaluation of the hyperparameters. Comparative analysis of (a) the kernel k of Optional near-far-aware Convolution $\mathcal{B}_\kappa$, and (b) the threshold for NALoss θ. (Unit: %)

NALoss. Our HACDR-Net is the baseline network. We surpass mainstream loss functions, achieving optimal outcomes (mAUPR:+1.85%, mIoU:+1.36%, mDice:+1.42%).

Visualization of Deep Features. T-SNE (Van der Maaten and Hinton 2008) is used to obtain 2D embeddings and visualize the deep features of the last encoder layer. As shown in Fig. 6, the lesion feature class generated by our network with NALoss is more compact, the difference between different classes is clearer, and the segmentation effect is improved.

Hyper-Parameters. We evaluate the influence of two core parameters on the model, one is the size of the convolution kernel k of ONFA branch $\mathcal{B}_\kappa$, and the other is the different threshold for NALoss θ, As in Fig. 7. ONFA branch adapts receptive field based on lesion size. The kernel k denotes convolution receptive field size. A larger kernel ($k = 17$) works well for EX segmentation, while a smaller kernel ($k = 5$) is good for MA. However, the best overall performance is achieved with a moderate k value ($k = 9$). Next, we examine the impact of various thresholds θ for NALoss. Setting θ to 0.5 yields the best results.

Conclusion

We introduce a new network, HACDR-Net, for DR multi-lesion segmentation. This network addresses the challenge of segmenting lesions of different shapes and sizes in DR images. Moreover, we propose a new loss function, NALoss, to handle imbalanced segmentation requirements. Our experiments show that HACDR-Net outperforms other methods in DR multi-lesion segmentation. In future work, we aim to enhance multi-lesion segmentation in DR images using multi-modal technology.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62301621, the Stable Support Projects for Shenzhen Higher Education Institutions under grant no. 20231122005530001, the Science and Technology Innovation Committee of Shenzhen Municipality under grant no. GJHZ20210705141812038.

References

Boyd, K.; Eng, K. H.; and Page, C. D. 2013. Area under the precision-recall curve: point estimates and confidence intervals. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*, 451–466.

Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; and Wang, M. 2022. Swin-unet: Unet-like pure transformer for medical image segmentation. In *European conference on computer vision*, 205–218.

Chen, J.; Lu, Y.; Yu, Q.; Luo, X.; Adeli, E.; Wang, Y.; Lu, L.; Yuille, A. L.; and Zhou, Y. 2021. Transunet: Transformers make strong encoders for medical image segmentation. *arXiv preprint arXiv:2102.04306*.

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

Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; and Shen, C. 2021. Twins: Revisiting the design of spatial attention in vision transformers. *Advances in Neural Information Processing Systems*, 34: 9355–9366.

Contributors, M. 2020. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark.

Cui, C.; Ren, Y.; Pu, J.; Li, J.; Pu, X.; Wu, T.; Shi, Y.; and He, L. 2023. A Novel Approach for Effective Multi-View Clustering with Information-Theoretic Perspective. *arXiv preprint arXiv:2309.13989*.

Diakogiannis, F. I.; Waldner, F.; Caccetta, P.; and Wu, C. 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162: 94–114.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Guo, M.-H.; Lu, C.-Z.; Liu, Z.-N.; Cheng, M.-M.; and Hu, S.-M. 2023. Visual attention network. *Computational Visual Media*, 1–20.

Guo, S.; Li, T.; Kang, H.; Li, N.; Zhang, Y.; and Wang, K. 2019. L-Seg: An end-to-end unified framework for multi-lesion segmentation of fundus images. *Neurocomputing*, 349: 52–63.

He, A.; Wang, K.; Li, T.; Bo, W.; Kang, H.; and Fu, H. 2022. Progressive Multiscale Consistent Network for Multiclass Fundus Lesion Segmentation. *IEEE Transactions on Medical Imaging*, 41(11): 3146–3157.

Huang, S.; Li, J.; Xiao, Y.; Shen, N.; and Xu, T. 2022. RTNet: relation transformer network for diabetic retinopathy multi-lesion segmentation. *IEEE Transactions on Medical Imaging*, 41(6): 1596–1607.

Li, T.; Gao, Y.; Wang, K.; Guo, S.; Liu, H.; and Kang, H. 2019. Diagnostic assessment of deep learning algorithms for diabetic retinopathy screening. *Information Sciences*, 501: 511–522.

Li, X.; Chen, H.; Qi, X.; Dou, Q.; Fu, C.-W.; and Heng, P.-A. 2018. H-DenseUNet: hybrid densely connected UNet for liver and tumor segmentation from CT volumes. *IEEE Transactions on Medical Imaging*, 37(12): 2663–2674.

Li, X.; Sun, X.; Meng, Y.; Liang, J.; Wu, F.; and Li, J. 2020. Dice Loss for Data-imbalanced NLP Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 465–476.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, 2980–2988.

Ling, Y.; Chen, J.; Ren, Y.; Pu, X.; Xu, J.; Zhu, X.; and He, L. 2023. Dual label-guided graph refinement for multi-view graph clustering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8791–8798.

Liu, C.; Wen, J.; Luo, X.; Huang, C.; Wu, Z.; and Xu, Y. 2023a. DICNet: Deep Instance-Level Contrastive Network for Double Incomplete Multi-View Multi-Label Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8807–8815.

Liu, C.; Wen, J.; Luo, X.; and Xu, Y. 2023b. Incomplete Multi-View Multi-Label Learning via Label-Guided Masked View- and Category-Aware Transformers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 8816–8824.

Liu, Q.; Liu, H.; Ke, W.; and Liang, Y. 2023c. Automated lesion segmentation in fundus images with many-to-many reassembly of features. *Pattern Recognition*, 136: 109191.

Liu, Z.; Hu, H.; Lin, Y.; Yao, Z.; Xie, Z.; Wei, Y.; Ning, J.; Cao, Y.; Zhang, Z.; Dong, L.; et al. 2022. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12009–12019.

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Milletari, F.; Navab, N.; and Ahmadi, S.-A. 2016. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, 565–571.

Porwal, P.; Pachade, S.; Kamble, R.; Kokare, M.; Deshmukh, G.; Sahasrabuddhe, V.; and Meriaudeau, F. 2018. Indian diabetic retinopathy image dataset (IDRiD): a database for diabetic retinopathy screening research. *Data*, 3(3): 25.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 658–666.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Shrivastava, A.; Gupta, A.; and Girshick, R. 2016. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 761–769.

Van der Maaten, L.; and Hinton, G. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. 2020. Deep high-resolution representation learning for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 43(10): 3349–3364.

Wang, Q.; Tao, Z.; Gao, Q.; and Jiao, L. 2022. Multi-view subspace clustering via structured multi-pathway network. *IEEE Transactions on Neural Networks and Learning Systems*.

Wang, Y.; Fei, J.; Wang, H.; Li, W.; Bao, T.; Wu, L.; Zhao, R.; and Shen, Y. 2023. Balancing Logit Variation for Long-tailed Semantic Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19561–19573.

Xie, S.; and Tu, Z. 2015. Holistically-nested edge detection. In *Proceedings of the IEEE international conference on computer vision*, 1395–1403.

Xu, J.; Tang, H.; Ren, Y.; Peng, L.; Zhu, X.; and He, L. 2022. Multi-level feature learning for contrastive multi-view clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16051–16060.

Yin, M.; Yao, Z.; Cao, Y.; Li, X.; Zhang, Z.; Lin, S.; and Hu, H. 2020. Disentangled non-local neural networks. In *Proceedings of the European conference on computer vision (ECCV)*, 191–207.

Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2881–2890.

Zhou, Y.; He, X.; Huang, L.; Liu, L.; Zhu, F.; Cui, S.; and Shao, L. 2019a. Collaborative Learning of Semi-Supervised Segmentation and Classification for Medical Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2019b. Unet++: Redesigning skip connections to exploit multiscale features in image segmentation. *IEEE Transactions on Medical Imaging*, 39(6): 1856–1867.