This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

CCT-Net: Category-Invariant Cross-Domain Transfer for Medical Single-to-Multiple Disease Diagnosis

Yi Zhou *[1], Lei Huang[2], Tao Zhou[3], and Ling Shao[4]

[1]School of Computer Science and Engineering, Southeast University, Nanjing, China
[2]SKLSDE, Institute of Artificial Intelligence, Beihang University, Beijing, China
[3]School of Computer Science and Technology, Nanjing University of Science and Technology, China
[4]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE

Abstract

A medical imaging model is usually explored for the diagnosis of a single disease. However, with the expanding demand for multi-disease diagnosis in clinical applications, multi-function solutions need to be investigated. Previous works proposed to either exploit different disease labels to conduct transfer learning through fine-tuning, or transfer knowledge across different domains with similar diseases. However, these methods still cannot address the real clinical challenge - a multi-disease model is required but annotations for each disease are not always available. In this paper, we introduce the task of transferring knowledge from single-disease diagnosis (source domain) to enhance multi-disease diagnosis (target domain). A category-invariant cross-domain transfer (CCT) method is proposed to address this single-to-multiple extension. First, for domain-specific task learning, we present a confidence weighted pooling (CWP) to obtain coarse heatmaps for different disease categories. Then, conditioned on these heatmaps, category-invariant feature refinement (CIFR) blocks are proposed to better localize discriminative semantic regions related to the corresponding diseases. The category-invariant characteristic enables transferability from the source domain to the target domain. We validate our method in two popular areas: extending diabetic retinopathy to identifying multiple ocular diseases, and extending glioma identification to the diagnosis of other brain tumors.

1. Introduction

Over the past decades, increasingly more automatic disease diagnosis systems have been developed for different medical imaging tasks [35, 51, 57, 59]. In some specific applications, large data with fine-grained annotations have facilitated model development and led to significant progress. For example, diabetic retinopathy (DR) diagnosis, including DR grading [19, 21] and lesion segmentation [44, 56], has been well studied on fundus images. Glioma identification, including tumor segmentation [20, 5, 9], has been successfully applied on brain MRI scans. Recently, many COVID-19 detection models [54, 27, 40] have also been rapidly developed using the annotated lung CT data. However, most of these works only focus on a single disease, which limits their transferability to other related diseases. In most clinical departments, multi-disease diagnosis (e.g. ocular diseases: DR, glaucoma, hypertension, myopia; brain tumors: glioma, meningioma, pituitary tumor; lung inflammation: community acquired, viral, bacteria pneumonia), has higher practical value and is preferred by clinicians. Therefore, as illustrated in Fig. 1, this work aims to transfer the ability of learning discriminative features from the well-

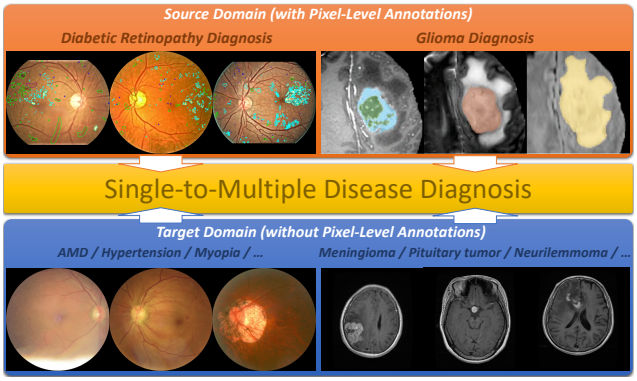

Figure 1. Motivation of the proposed CCT-Net. The knowledge learned from the well-explored single-disease diagnosis with fine-grained annotations can be transferred to improve the diagnosis of multiple related diseases without annotations.

*Corresponding author: Yi Zhou (yizhou.szcn@gmail.com)

explored single-disease diagnosis with rich annotations to multi-disease diagnosis with limited annotations.

Previous works on transfer learning from one disease to another can be coarsely categorized into two types. The first type adopts models pre-trained on annotated natural images [29] or similar-domain data and then fine-tune them for new diseases [26, 19]. The low-level features of similar-domain data can be shared so that the model only needs to transfer the high-level semantic disease category information. However, this approach usually requires extensive annotation for new diseases. Since annotating medical data requires professional knowledge and is time-consuming, we cannot expect rich annotations to be available for all diseases. The second type focuses on domain adaptation [49, 12, 4, 30, 15, 34] for different-domain data that have the same disease or lesion types. These works aim to transfer domain-invariant knowledge from labeled source domain data to unlabeled target domain data. For example, unsupervised lesion segmentation models for endoscopic data [11] have been explored through domain adaptation using labeled gastroscope segmentation data. However, this transfer is only feasible when the disease categories of the two domains are the same: cancer, polyp, gastritis, ulcer and bleeding. In more general clinical scenarios, multi-disease diagnosis of diseases that do not share similar appearances is required, but annotations for each disease are not always available. Thus, it is difficult to directly apply these existing models to address the practical problems.

In this paper, we focus on the problem where the two domains have different disease identification tasks, with the source domain having pixel-level labels for a single disease and the target domain involving more diseases, without fine-grained annotations (only image-level disease categories are known). We propose a category-invariant cross-domain transfer method to learn the knowledge from the source domain task and improve both the classification and localization performance of the target domain task. **The main contributions of this work are as follows:**

1. A domain-specific task learning module is designed to learn domain-invariant features while preserving the disease discrepancy between two domains. A CWP global pooling method is proposed to obtain better class activation maps (CAMs) than other global pooling operations.

2. Conditioned on the coarse heatmaps of different categories, we propose CIFR blocks to construct the CCT-Net for localizing more discriminative regions of the corresponding diseases. The category-invariant characteristic enables the transferability from the source domain to target domain. Moreover, such refined features can contribute to the final classification performance as well.

3. Experimental evaluations are conducted in two popular medical imaging tasks. First, we extend DR diagnosis on fundus images to the diagnosis of multiple ocular diseases, such as glaucoma and hypertension. Second, glioma segmentation on brain MRI scans is exploited to improve the segmentation and classification performance of other tumors, such as meningioma and pituitary tumors. Experimental results demonstrate the effectiveness of our method.

2. Related Work

2.1. Medical Disease Diagnosis Scenarios

Deep neural networks have achieved significant success in the diagnosis of numerous individual diseases from medical imaging data. For example, in fundus imaging for ocular diseases, DR and glaucoma have been widely explored, with tasks including DR grading [19], DR lesion semantic segmentation [14], and glaucoma detection [16]. In chest X-rays for thoracic diseases, pneumonia [45] and tuberculosis [36] identification models have been developed and used in clinical applications. In brain MRI, researchers are most interested in glioma segmentation, and the BraTS [39] competition is organized annually to provide a platform for contributing to the community. Moreover, in lung CT, many well-developed lung nodule [52] and pneumonia [13] detection systems have achieved satisfactory performance and reached radiologist level. However, these single-disease diagnosis models have limited transferability to multiple diseases and usually require new annotations. Multi-disease diagnosis systems are more practical.

2.2. Cross Domain Transfer Learning

Domain adaptation (DA) [49] is a way of transfer learning which deals with scenarios in which a model trained on a source distribution is used in the context of a different (but related) target distribution. DA methods aim to learn domain-invariant representations to address domain shifts. Adversarial networks [23] and diverse variants [10, 42, 33] based on the adversarial strategy have been widely explored for domain alignment. For example, the domain-adversarial neural network (DANN) [17] introduces a confusion loss to match the distributions of the source and target domains in order to confuse the high-level classification layers. Meanwhile, maximum classifier discrepancy (MCD) [47] was presented to utilize task-specific decision boundaries to align distributions. Except for the adversarial approach, divergence-based DA methods [53, 48] aim to minimize the divergence criterion between the source and target domains, while batch normalization [32, 3] parameters have been used to model domain-specific information. DA has also been addressed by adopting auxiliary reconstruction tasks [18, 61] to create a shared representation for each of the domains. Most of these methods define DA to be a problem in which the task space is similar to the source space, with the only difference being the input domain divergence. However, disease discrepancy exists in our task.

3. Proposed Methods

3.1. Problem Formulation

In the proposed domain transfer scenario, we are given the source domain data $\mathcal{X}_S$ with pixel-level labels $\mathcal{Y}_S^p$ and the target domain data $\mathcal{X}_T$ without pixel-level labels. The disease categories of two domains are denoted as $\mathcal{Y}_S^c$ and $\mathcal{Y}_T^c$, respectively. $\mathcal{Y}_S^p$ and $\mathcal{Y}_S^c$ have the same category cardinality N_S, while $\mathcal{Y}_T^c$ has N_T categories. The overall aim of the task is to transfer knowledge from the source domain data ($\mathcal{X}_S$, $\mathcal{Y}_S^p$, $\mathcal{Y}_S^c$) to improve both the localization and classification performance on the target domain data ($\mathcal{X}_T$, $\mathcal{Y}_T^c$).

As illustrated in Fig. 2, the domain-specific encoder (DSE), marked in green and purple, is designed for learning different disease classification tasks in the two domains and optimized by the corresponding classification loss $\mathcal{L}_{Cls}$. The coarse heatmaps can be obtained to weight the features for further learning discriminative feature refinement. Then, the category-invariant feature refinement module $f_{CIFR}(\cdot)$ is introduced to decode different category heatmaps to the corresponding pixel-level probabilistic predictions. Moreover, the features refined by the CIFR can be further adopted to improve the performance of classification tasks. The overall optimization function is formulated as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{L}_{Cls}^S(\mathcal{X}_S, \mathcal{Y}_S^c) + \mathcal{L}_{Cls}^T(\mathcal{X}_T, \mathcal{Y}_T^c) + \lambda \mathcal{L}_{FR}(\mathcal{X}_S, \mathcal{Y}_S^p) \\ &= \ell(f_{DSE}(\mathcal{X}_S), \mathcal{Y}_S^c) + \ell(f_{DSE}(\mathcal{X}_T), \mathcal{Y}_T^c) + \\ &+ \lambda \ell(f_{CIFR}(f_{DSE-1}(\mathcal{X}_S)), \mathcal{Y}_S^p),\end{aligned} \tag{1}$$

where λ balances the weights of $\mathcal{L}_{Cls}$ and the feature refinement $\mathcal{L}_{FR}$ loss, and $\ell(\cdot)$ denotes the cross-entropy loss.

3.2. Domain-Specific Task Learning

The basic disease classification frameworks used in most previous works [19, 38, 45] typically adopt a classic deep neural network (e.g. ResNet [22] or DenseNet [24]) and optimize it with disease category information. Since image-level disease category labels are easy to collect, a baseline classification result and coarse localization map can be obtained. In this work, we aim to optimize a DSE for learning the source and target domain tasks simultaneously. Both domains share convolutional ($Conv$) parameters to constrain the encoder to learn domain-invariant representations. Since the two domains are usually from different data sources and their disease categories have discrepancies, both the low-level and high-level features have distribution differences between the two domains. Thus, in addition to exploiting separate task-specific classification losses to preserve the disease discrepancy, we adopt a domain-specific batch normalization (BN) inspired by [3] to enhance the learning of domain-invariant features.

To construct the DSE, DenseNet-121 [24] is adopted as the backbone. The BN layer after each $Conv$ layer for different domains is learned separately. Domain-specific affine parameters are allocated to estimate different batch statistics for each domain. We expect the DSE to be able to learn domain-invariant representations because the domain-specific input information within the network can be effectively removed via the captured statistics and learned parameters from the given domain. Note that the better the domain-invariant features extracted by the DSE, the more effectively the feature refinement module trained on the source domain data can be transferred to the target domain.

3.2.1 Confidence Weighted Pooling

In addition to obtaining a preliminary classification result of different diseases, the DSE is also able to compute coarse heatmaps for different categories. In the image classification task, weakly supervised methods [8, 50] using only image-level labels usually adopt a class activation map (CAM) [55] to compute a coarse heatmap for localizing each category. The CAM for a particular category indicates the discriminative image regions used by the network to identify that category and can be used to interpret the prediction decision made by the network. This localization ability is enabled by the basic global average/max pooling (GAP/GMP) layer. Although GAP and GMP have been widely used, they are not trainable and the CAM may fail to localize the most discriminative regions. Log-Sum-Exp (LSE) pooling [43] introduces a hyper-parameter γ to serve as an adjustable option between max pooling and average pooling. However, LSE pooling is still not optimizable and has overflow or underflow problems.

In this work, we propose a trainable CWP global pooling method to enhance the original CAM with better localization performance. Given the last-layer feature maps of the DSE, the input image is represented as $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, where N denotes the number of image regions and $\mathbf{x}_n$ is the feature embedding of the n-th region. For each category, the $\{\mathbf{x}_n\}_{n=1}^N$ are passed forward to a 1×1 $Conv$ layer to compute the class confidence score, which indicates the disease possibility of each embedding. Rather than directly adopting the confidence score map as the localization map, we further employ a sigmoid function $\sigma(\cdot)$ and normalize the activation map to globally pool the $\{\mathbf{x}_n\}_{n=1}^N$ by normalized confidence weights. Then, during training, the pooled embedding is passed forward to the same 1×1 $Conv$ layer to learn image-level disease classification. In the test phase, the probability map after $\sigma(\cdot)$ is adopted as the localization map. Overall, the CWP is defined as:

$$\mathbf{x} = \sum_{n=1}^{N} \frac{\sigma(\mathbf{w}\mathbf{x}_n + b)}{\sum_{n=1}^{N} \sigma(\mathbf{w}\mathbf{x}_n + b)} \mathbf{x}_n, \tag{2}$$

where $\mathbf{w}$ and b are the 1×1 $Conv$ parameters for learning classifiers. The heatmaps marked in gray in Fig. 2 are examples obtained by CWP-based CAM. CWP is used for

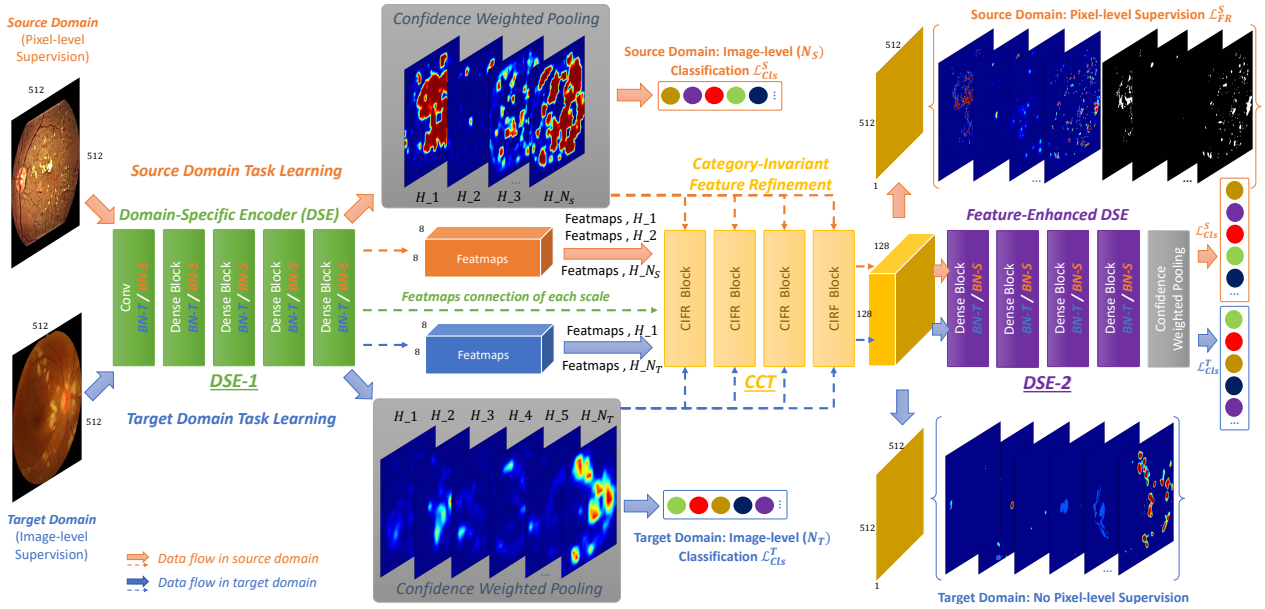

Figure 2. Pipeline of the proposed method. The source domain data has pixel-level labels, while the target domain data only has image-level labels. The DSE in green (DSE-1) is proposed to obtain preliminary classification results and category-specific coarse localization maps(H_n). The CCT-Net enables feature refinement and transfers such ability learned from the source domain, through CIFR blocks, to implement segmentation for disease categories in the target domain. The enhanced features, focusing on more discriminative regions, are further encoded by the domain-specific encoder in purple (DSE-2) to improve the disease classification performance.

both the source and target domain task learning. Although the initial CWP-based CAM already discards large irrelevant regions to discriminate correct diseases, its coarse heatmaps still contain some distracting information from noise or misclassified categories. In this work, we aim to refine these coarse heatmaps to fine-grained pixel-level predictions that delineate more discriminative regions, and also transfer this refinement ability from the source to the target domain. Please note that CAM might not be the best weakly-supervised localization method [8], but that is not the primary concern of this work. We will investigate this point specifically to enhance the CCT-Net in future work.

3.3. Category-Invariant Cross-Domain Transfer

Given images with pixel-level annotations, semantic segmentation models [31, 13] are usually adopted to obtain fine-grained predictions. However, in our task, only the source domain $\mathcal{X}_S$ is provided with pixel-level $\mathcal{Y}_S^p$, while the target domain $\mathcal{X}_T$ is unlabeled and contains different categories. Thus, we propose the CCT-Net to transfer the feature refinement ability learned from the source domain to enable pixel-level segmentation for new categories in the target domain. Moreover, the refined features can be further used for improving the classification performance.

3.3.1 Category-Invariant Feature Refinement

The CCT-Net, stacking multiple sets of CIFR blocks, inherits some characteristics from the expansive module of conventional image segmentation frameworks [46, 60] but has essential differences. As shown in the top-left of Fig. 3, to perform pixel-level segmentation, the standard expansive path in segmentation networks usually decodes the bottleneck features of the input image into C-channel vectors for C-category output masks. Essentially, the expansive module takes one input and learns a C-channel classifier at the end. However, for CCT-Net, category-specific coarse heatmaps of different categories predicted by the DSE are combined with the bottleneck features to construct multiple inputs. The pipeline of CCT-Net is shown in the top-right of Fig. 3. For each input tuple (*i.e.* bottleneck features + one category heatmap), CCT-Net refines the category-specific features and predicts a one-channel output vector supervised by the category's ground truth mask. Note that CCT-Net plays the role of classifying whether or not a pixel belongs to a specific category conditioned on that category's coarse heatmap, rather than discriminating which category a pixel should be. Thus, adopting tuples of different categories to train CCT-Net makes it category-invariant for feature refinement. Moreover, since the DSE is able to extract domain-invariant features, CCT-Net can effectively transfer the refinement ability learned from the source domain to all categories in the target domain.

In this work, the CCT-Net consists of four CIFR blocks whose operations are shown in the bottom of Fig. 3. In a block of a certain scale, different category-specific coarse heatmaps are first separately concatenated with the input

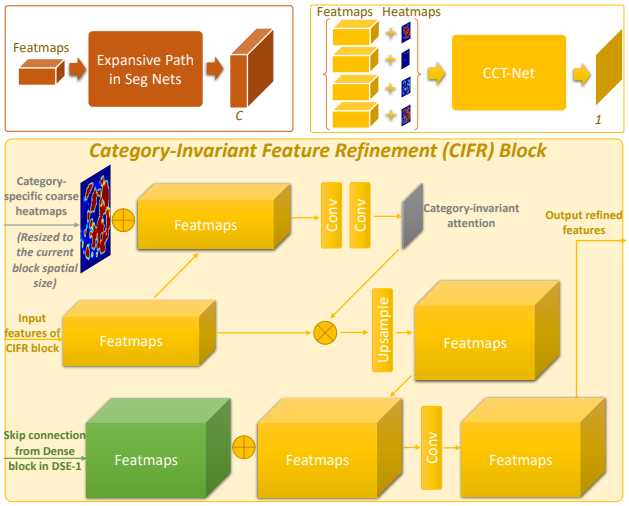

Figure 3. Upper part: pipeline comparison of the standard expansive path of segmentation networks [46] and the proposed CCT-Net. Bottom part: structural details of the CIFR Block.

feature maps. The heatmaps obtained by CWP-based CAM of the DSE are resized to fit the spatial size of the current block. Next, two 1×1 $Conv$ layers are added to reduce the channel dimension for computing the attention mask $\mathcal{A} \in \mathbb{R}^{H \times W}$, where H and W denote the width and height, respectively. Then, $\mathcal{A}$ is normalized using the softmax function to obtain the final category-invariant attention α as:

$$\alpha(x, y) = \frac{e^{\mathcal{A}(x,y)}}{\sum_{i=1}^{H} \sum_{j=1}^{W} e^{\mathcal{A}(i,j)}}, \tag{3}$$

where $\alpha(i, j)$ denotes the value of the category-invariant attention α at (i, j). We adopt α to weigh the input feature maps and then conduct upsampling through bilinear interpolation. All the remaining operations of the CIFR block are similar to the expansive path in [46]. A skip connection from the same-scale dense block of the DSE is used for concatenation, which can recover spatial information lost during downsampling. Finally, a 3×3 $Conv$ layer is used to compute the refined output feature maps. Please note that the network parameters of the CCT-Net are only optimizable when training on the source domain data with pixel-level supervision, and fixed for the target domain data flow.

3.3.2 Feature-Enhanced Domain Task Learning

As mentioned, besides providing the pixel-level segmentation performance, CCT-Net essentially refines the feature representations to delineate more discriminative regions. As shown in purple in Fig. 2, we employ the same network architecture as the DSE-1, except the first $Conv$ layer, to build a feature-enhanced DSE (DSE-2) to further improve disease classification performance. We conduct element-wise max on the last-layer feature maps of the CCT-Net along all the categories and take the results as inputs of the DSE-2.

3.4. Implementation Details

Although our training scheme can adopt an end-to-end strategy from the beginning, we observe better optimization results when we pre-train the DSE-1 first and then add the CCT-Net and feature-enhanced DSE-2 for combined training. This is because training the CCT-Net requires effective category-specific heatmaps to be provided in the very first stage. Moreover, the CCT-Net is only trained on the source domain data flow due to its pixel-level supervision. Parameter λ in Eq. 1 is set to 0.5 throughout our experiments, which yields the best performance. Other hyper-parameters are set as follows. The Adam optimizer is adopted with an initial learning rate of 0.001 and default parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$. The mini-batch size is set to 32.

4. Experiments and Results

4.1. Datasets and Evaluation Metrics

Ocular Diseases (Fundus Imaging) - FGADR [58] is a fine-grained annotated DR dataset with 1,842 pixel-level labeled images. The DR related lesions, including microaneurysms (MA), hard exudates (EX), soft exudates (SE), hemorrhages (HE), intra-retinal microvascular abnormalities (IRMA), and neovascularization (NV), are annotated by three ophthalmologists. **ODIR-5K** [1] consists of 7,000 fundus images with multi-label image-level annotations. The labels contain eight ocular disease categories, including diabetes, glaucoma, cataract, age-related macular degeneration (AMD), hypertension, myopia, normal, and other diseases. Therefore, we adopt the FGADR and ODIR-5K datasets as the source and target domain data, respectively, in our task. Moreover, to obtain image-level labels for the FGADR dataset, if the ground-truth mask of a category has annotated spots, the corresponding image-level label is marked as positive, and negative otherwise. The **evaluation** of our task consists of two steps. The first step is to evaluate the performance of the category-invariant feature refinement module at a pixel level (same as segmentation) using the dice score (Dice) and area under the curve of precision-recall (AUC-PR). The second step is to evaluate the multi-ocular-disease identification results at an image level using the Cohen's kappa, F-1 score, and AUC of receiver operating characteristic (ROC).

Brain Diseases (MRI) - BraTS 2019 [39] provides pre-operative multimodal MRI scans from 335 patients and focuses on the segmentation of intrinsically heterogeneous brain tumors, namely glioma. Each patient case contains 155 slices. The pixel-level annotations comprise GD-enhancing tumors (ET), peritumoral edema (ED), and necrotic and non-enhancing tumor cores (NET). We use BraTS 2019 as the source domain data. **BrainTumor** [7] has 3,064 T1-weighted contrast-enhanced (T1-ce) images from 233 patients with three kinds of brain tumor: menin-

gioma, glioma, and pituitary tumor. Both pixel-level and image-level labels are provided. We use BrainTumor as the target domain data, so its pixel-level annotations are only used for testing. Please note that since BrainTumor only has the T1-ce modality, we also only adopt the T1-ce modality images of BraTS 2019 for consistency. The **evaluation metrics** for pixel-level segmentation are Dice and Hausdorff (Haus.) distance, while accuracy and confusion matrix are adopted to evaluate image-level classification.

4.2. Single-to-Multiple Ocular Disease Diagnosis

DR is the most widely studied eye disease, which can damage the blood vessels in the back of the retina and lead to blindness. In addition to DR, other ocular diseases, such as glaucoma, cataracts, AMD, and hypertension, are also important but have less research data. In this experiment, we assess the CCT-Net's ability to transfer the knowledge learned from DR data to improve multiple ocular disease diagnosis. Different pixel-level annotated lesion categories, including MA, HE, SE, and EX from the FGADR dataset, are used to train the CCT-Net. Their corresponding image-level labels are used to train the DSE for the source domain. Multi-disease image-level labels from the ODIR-5K dataset are used to train the DSE for the target domain flow. We study the effectiveness of our method in terms of both pixel-level segmentation and image-level classification.

4.2.1 Evaluation of Pixel-Level Mask Segmentation

Our primary concern is to investigate the effectiveness of the cross-domain category-invariant feature refinement of the CCT-Net. We first validate this point by evaluating the pixel-level segmentation performance and comparing the masks refined by the CCT-Net with standard segmentation networks. In this ocular task, since the pixel-level ground truths of the target domain data are not available, we report the segmentation results for the source domain in Table 1. However, a qualitative visualization of pixel-level segmentation for both the source and target domain is provided in Fig. 4, which also effectively proves the transferability. The quantitative results for evaluating the transferability to the target domain will be given in the next task of brain MRI.

Multiple baselines and state-of-the-art segmentation models are compared with our method in Table 1. First, **w/o-CCT** does not apply CCT but directly resizes the coarse map predictions (binarized by the threshold 0.2) obtained by the CWP-based DSE-1 to the size of the ground truths for evaluation. Moreover, to explore how the coarse localization performance produced by different global pooling operations affects the CCT-Net, the masks predicted (binarized by the threshold 0.2) by our **CCT-w-CWP** for segmentation are compared with those produced by three other baselines: **CCT-w-AVG**, **CCT-w-MAX**, and **CCT-w-LSE**. Four other traditional segmentation networks, including FCN-8s [37], DeepLab_v3+ [6] (s=8), U-Net [46], and Attention U-Net [41], are also compared. Following [58], a two-fold cross validation is adopted.

Table 1. Ocular: segmentation performance in source domain. The two best results are in **red** and blue. **'w/o'='without', 'w'='with'.**

Methods	MA		HE		SE		EX	
	Dice	PR	Dice	PR	Dice	PR	Dice	PR
FCN-8s [37]	0.468	0.363	0.509	0.606	0.637	0.642	0.586	0.686
DL_V3+ [6]	0.482	0.364	0.550	0.619	0.648	0.659	0.602	0.702
U-Net [46]	0.521	0.382	0.570	0.643	0.655	0.683	0.607	0.726
Att. U-Net [41]	0.536	0.435	0.576	0.678	0.689	0.712	0.637	0.762
w/o-CCT	0.453	0.359	0.471	0.554	0.633	0.638	0.569	0.653
CCT-w-AVG	0.493	0.369	0.553	0.622	0.651	0.658	0.599	0.705
CCT-w-MAX	0.491	0.368	0.552	0.622	0.649	0.655	0.596	0.703
CCT-w-LSE	0.517	0.383	0.567	0.640	0.663	0.681	0.616	0.723
CCT-w-CWP	**0.542**	**0.452**	**0.591**	**0.687**	**0.709**	**0.721**	**0.644**	**0.768**

Table 2. Ocular: classification results in target domain.

	Methods	Kappa	F-1	AUC
Baseline	DSE-1-w-AVG	0.6556	0.9163	0.9274
	DSE-1-w-MAX	0.6548	0.9155	0.9271
	DSE-1-w-LSE	0.6561	0.9168	0.9279
	DSE-1-w-CWP	0.6712	0.9195	0.9298
Transfer	DANN [17]	0.6934	0.9278	0.9381
	MCD [47]	0.7296	0.9386	0.9477
	ITL [58]	0.7348	0.9426	0.9498
	DSE-1+CCT*+DSE-2	0.6747	0.9183	0.9309
	DSE-1+CCT+DSE-2_B1	0.7253	0.9369	0.9448
	DSE-1+CCT+DSE-2_B2	**0.7508**	**0.9510**	**0.9583**

As shown in Table 1, we observe that the coarse heatmaps obtained by DSE-1 using different global pooling operations have different effects on the performance of the CCT-Net. The CCT-w-AVG and CCT-w-MAX achieve similar results but perform worse than the U-Net, since these two basic global pooling methods do not have good enough preliminary localization performance. Compared to CCT-w-AVG, CCT-w-LSE increases the Dice and AUC of PR by, on average, 1.7% and 1.8%, respectively, while the performance is significantly improved by CCT-w-CWP, with increases of 4.75% and 6.95% in terms of Dice and PR. This illustrates that the CWP-based DSE-1 obtains better localization performance and benefits the training of the CCT-Net. Once the CCT-Net is detached (*i.e.* w/o-CCT), the segmentation performance using coarse maps is poor. Moreover, although traditional segmentation networks tackle the segmentation task as a classification problem on each pixel using C channels for the output classifiers, our category-invariant feature refinement mechanism can also achieve competitive segmentation performance to these, with even a slight improvement. Fig. 4 shows various lesion regions related to multiple diseases in target domain are segmented.

4.2.2 Evaluation of Image-Level Disease Classification

Although the target domain data does not have pixel-level annotations, we validate whether the feature refinement ability of the CCT-Net can be transferred from the source domain to the target domain for improving the multi-disease classification performance. We report the classification re-

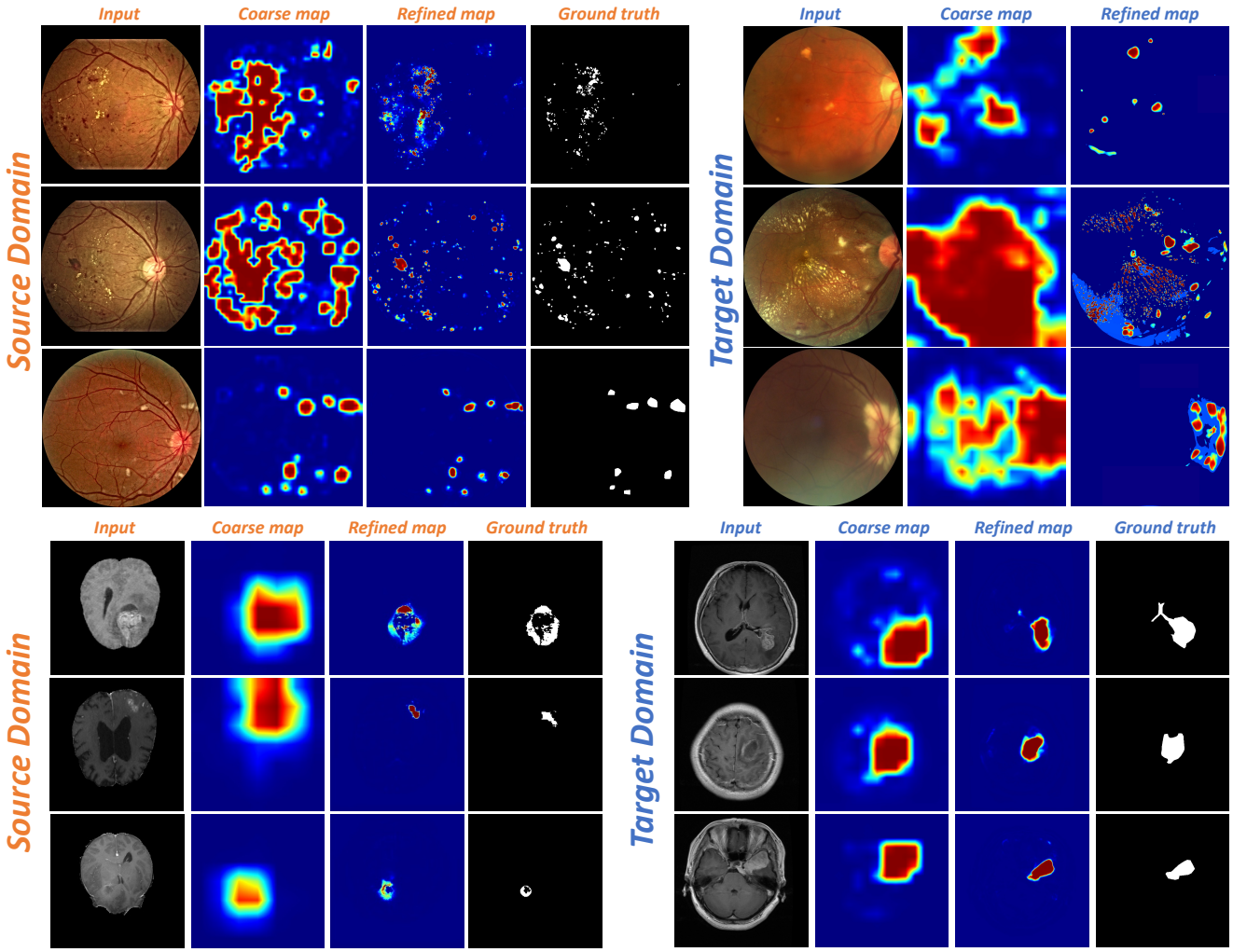

Figure 4. Qualitative results visualization of the pixel-level mask segmentation. Each row shows different disease categories.

sults for the target domain in Table 2. First, DSE-1 models with different global pooling operations, including **DSE-1-w-AVG**, **DSE-1-w-MAX**, **DSE-1-w-LSE**, and **DSE-1-w-CWP**, are compared to provide baseline results without using CCT. Moreover, to study the effectiveness of CCT, **DSE-1+CCT+DSE-2_B1** and **DSE-1+CCT+DSE-2_B2** are explored, which adopt the standard BN and domain-specific BN in the DSE, respectively. To determine whether the performance improvements arise from the novel model designs or the increase in parameters, we also compare **DSE-1+CCT*+DSE-2**, which drops the supervision of CCT but still keeps the network in the pipeline. The CWP global pooling is employed in the DSE of these three methods. We also compare our model with three state-of-the-art transfer learning methods **DANN** [17], **MCD** [47], and **ITL** [58]. In MCD, pixel-level annotated data is added to optimize the discrepancy loss for the source domain. Following [58], the training and testing set is split as 4:1 for five-fold cross validation.

As shown in Table 2, for the baselines using DSE without the CCT-Net, the CWP global pooling performs better than other pooling methods. The Kappa of DSE-1-w-CWP is slightly increased by 1.56% compared to that of DSE-1-w-AVG. The transfer learning methods MCD and ITL, which exploit the knowledge learned from the pixel-level annotated source domain data, significantly increase the Kappa by 5.84% and 6.36%, respectively, compared to DSE-1-w-CWP. Our DSE-1+CCT+DSE-2_B1, adopting the features refined by the CCT-Net, also improves the classification results in the target domain, with a Kappa increase of 5.41%. Since there exists a difference in distribution between the source and target domain data, enhancing the domain-invariant feature extraction ability of the DSE can make the CCT-Net perform category-invariant feature refinement for the target domain better. We adopt the domain-specific BN in the DSE, denoted as DSE-1+CCT+DSE-2_B2, to further increase the Kappa by 2.55%. DSE-1+CCT*+DSE-2 illustrates that CCT is not effective when dropping the $\mathcal{L}_{FR}$.

Table 3. Brain: segmentation performance in two domains. The two best results are in **red** and blue. **'w/o'='without', 'w'='with'.**

		Methods	ET		ED		NET	
			Dice	Haus.	Dice	Haus.	Dice	Haus.
Source Domain	Supervised	FCN-8s [37]	0.731	5.87	0.485	17.48	0.650	5.83
		DL_V3+ [6]	0.747	5.70	0.500	17.02	0.669	5.23
		U-Net [46]	0.762	5.46	0.514	16.73	0.680	4.79
		Att. U-Net [41]	0.775	5.23	0.528	16.25	0.701	4.52
		w/o-CCT	0.608	9.71	0.424	20.18	0.604	9.83
		CCT-w-AVG	0.745	5.67	0.499	17.03	0.669	5.18
		CCT-w-MAX	0.743	5.88	0.497	16.94	0.661	5.31
		CCT-w-LSE	0.763	5.42	0.521	16.56	0.683	4.71
		CCT-w-CWP	**0.786**	**5.08**	**0.541**	**15.96**	**0.714**	**4.38**

		Methods	Meningioma		Glioma		Pituitary Tumor	
			Dice	Haus.	Dice	Haus.	Dice	Haus.
Target Domain	Supervised	FCN-8s [37]	0.824	6.17	0.621	11.30	0.785	5.09
		DL_V3+ [6]	0.851	5.65	0.638	10.73	0.812	4.73
		U-Net [46]	0.875	5.20	0.647	10.61	0.830	4.49
		Att. U-Net [41]	**0.889**	**4.97**	0.665	10.28	**0.849**	**4.15**
	WeaklySupervised	w/o-CCT	0.715	8.55	0.531	15.26	0.660	6.93
		CCT-w-AVG	0.845	5.76	0.629	10.99	0.793	4.95
		CCT-w-MAX	0.838	5.88	0.624	11.18	0.792	4.95
		CCT-w-LSE	0.862	5.38	0.641	10.76	0.819	4.62
		CCT-w-CWP_B1	0.873	5.21	0.653	10.55	0.830	4.46
		CCT-w-CWP_B2	0.883	5.05	**0.668**	**10.22**	0.841	4.27

4.3. Single-to-Multiple Brain Tumor Identification

A brain tumor is a mass or growth of abnormal cells in the brain. Many different types of brain tumors can be identified through brain MRI. Some brain tumors are benign, while others are malignant. Glioma is the most frequent primary brain tumor, which originates from glial cells and infiltrates the surrounding tissues. Besides, meningioma arises from the membranes that surround the brain and spinal cord, and pituitary tumors are abnormal growths that develop in the pituitary gland. In this experiment, glioma data with pixel-level labels of ET, ED, and NET, provided by BraTS 2019, are used to train the CCT-Net. The classification task in the target domain is to discriminate between the three types of tumors, meningioma, pituitary tumor, and glioma, from the BrainTumor dataset.

4.3.1 Evaluation of Pixel-Level Mask Segmentation

In this task, both the source and target domains have pixel-level ground truths, but we still only adopt the source domain annotations to train the CCT-Net and the target domain annotations for testing. Thus, for the target domain, our CCT-Net for segmentation is weakly supervised, which can better illustrate the cross-domain transferability of our method. For both the source and target domains, five-fold cross validation based on patient IDs is adopted. In the target domain data flow, only the image-level category labels of the training set are used. As shown in the target domain results of Table 3, our weakly-supervised CCT-based methods can achieve competitive performance compared to the traditional supervised segmentation networks trained using pixel-level labels. **CCT-w-CWP_B1**, which adopts the standard BN in the DSE, already slightly outperforms some segmentation models, validating the category-invariant feature refinement ability of the CCT-Net. Moreover, **CCT-w-CWP_B2**, which uses the domain-specific BN in the DSE to learn better domain-invariant representations, further improves the CCT in terms of cross-domain transferability. Some qualitative results are visualized in the bottom part of Fig. 4. We also compare our method to some weakly-supervised models [28, 25, 2] in the supplementary file.

Table 4. Brain: classification accuracy in target domain. The two best results are in **red** and blue. **'w'='with'.**

	Methods	Meningioma	Glioma	Pituitary Tumor
Baseline	DSE-1-w-AVG	0.8863	0.9194	0.9275
	DSE-1-w-MAX	0.8845	0.9205	0.9282
	DSE-1-w-LSE	0.9048	0.9379	0.9441
	DSE-1-w-CWP	0.9067	0.9394	0.9438
Transfer	DANN [17]	0.9092	0.9409	0.9472
	MCD [47]	0.9123	0.9446	0.9531
	ITL [58]	0.9159	0.9472	0.9555
	DSE-1+CCT*+DSE-2	0.9061	0.9389	0.9442
	DSE-1+CCT+DSE-2_B1	0.9246	0.9503	0.9560
	DSE-1+CCT+DSE-2_B2	**0.9379**	**0.9568**	**0.9614**

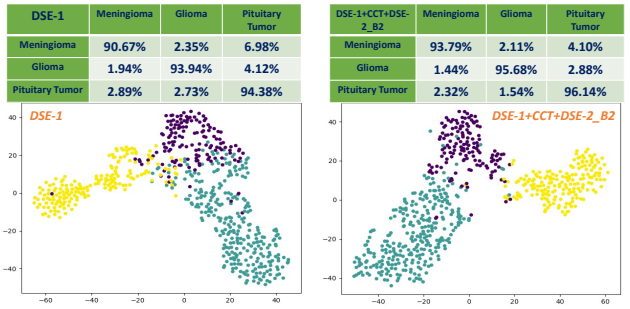

Figure 5. Comparison of confusion matrix and feature visualization for DSE-1 and DSE-1+CCT+DSE-2_B2.

4.3.2 Evaluation of Image-Level Disease Classification

The training and testing split in this study is the same as the split for pixel-level mask segmentation. Since the target domain task is single-label disease category classification, the classification accuracies are shown in Table 4. The average accuracy for the three brain tumors is increased by our fully-equipped DSE-1+CCT+DSE-2_B2, with an increase of 2.21% compared to DSE-1-w-CWP. To better demonstrate the detailed improvements by the CCT-Net, we also visualize the discriminative feature space and the confusion matrix of the two compared methods in Fig. 5.

5. Conclusion

We propose to extend single-disease to multi-disease diagnosis to better serve clinical needs. The CCT-Net is designed for category-invariant cross-domain transfer to learn knowledge from source domain to improve the results of target domain. We validated our method in two popular medical imaging areas. This work was supported by the National Natural Science Foundation of China (62106043).

References

[1] International competition on ocular disease intelligent recognition. https://odir2019.grand-challenge.org.

[2] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In *CVPR*, pages 2209–2218, 2019.

[3] Woong-Gi Chang, Tackgeun You, Seonguk Seo, Suha Kwak, and Bohyung Han. Domain-specific batch normalization for unsupervised domain adaptation. In *CVPR*, pages 7354–7362, 2019.

[4] Cheng Chen, Qi Dou, Hao Chen, Jing Qin, and Pheng-Ann Heng. Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In *AAAI*, volume 33, pages 865–872, 2019.

[5] Chen Chen, Xiaopeng Liu, Meng Ding, Junfeng Zheng, and Jiangyun Li. 3d dilated multi-fiber network for real-time brain tumor segmentation in mri. In *MICCAI*, pages 184–192. Springer, 2019.

[6] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, pages 801–818, 2018.

[7] Jun Cheng, Wei Huang, Shuangliang Cao, Ru Yang, Wei Yang, Zhaoqiang Yun, Zhijian Wang, and Qianjin Feng. Enhanced performance of brain tumor classification via tumor region augmentation and partition. *PloS one*, 10(10):e0140381, 2015.

[8] Junsuk Choe, Seong Joon Oh, Seungho Lee, Sanghyuk Chun, Zeynep Akata, and Hyunjung Shim. Evaluating weakly supervised object localization methods right. In *CVPR*, pages 3133–3142, 2020.

[9] Jose Dolz, Karthik Gopinath, Jing Yuan, Herve Lombaert, Christian Desrosiers, and Ismail Ben Ayed. Hyperdense-net: a hyper-densely connected cnn for multi-modal image segmentation. *IEEE Transactions on Medical Imaging*, 38(5):1116–1126, 2018.

[10] Jiahua Dong, Yang Cong, Gan Sun, and Dongdong Hou. Semantic-transferable weakly-supervised endoscopic lesions segmentation. In *ICCV*, pages 10712–10721, 2019.

[11] Jiahua Dong, Yang Cong, Gan Sun, Bineng Zhong, and Xiaowei Xu. What can be transferred: Unsupervised domain adaptation for endoscopic lesions segmentation. In *CVPR*, pages 4023–4032, 2020.

[12] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of convnets for biomedical image segmentations with adversarial loss. In *IJCAI*, pages 691–697, 2018.

[13] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 2020.

[14] Alex Foo, Wynne Hsu, Mong-Li Lee, Gilbert Lim, and Tien Yin Wong. Multi-task learning for diabetic retinopathy grading and lesion segmentation. In *AAAI*, pages 13267–13272, 2020.

[15] Bo Fu, Zhangjie Cao, Mingsheng Long, and Jianmin Wang. Learning to detect open classes for universal domain adaptation. In *ECCV*. Springer, 2020.

[16] Huazhu Fu, Jun Cheng, Yanwu Xu, Damon Wing Kee Wong, Jiang Liu, and Xiaochun Cao. Joint optic disc and cup segmentation based on multi-label deep network and polar transformation. *IEEE Transactions on Medical Imaging*, 37(7):1597–1605, 2018.

[17] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030, 2016.

[18] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, David Balduzzi, and Wen Li. Deep reconstruction-classification networks for unsupervised domain adaptation. In *ECCV*, pages 597–613. Springer, 2016.

[19] Varun Gulshan, Lily Peng, Marc Coram, Martin C Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *Jama*, 316(22):2402–2410, 2016.

[20] Mohammad Havaei, Axel Davy, David Warde-Farley, Antoine Biard, Aaron Courville, Yoshua Bengio, Chris Pal, Pierre-Marc Jodoin, and Hugo Larochelle. Brain tumor segmentation with deep neural networks. *Medical Image Analysis*, 35:18–31, 2017.

[21] Along He, Tao Li, Ning Li, Kai Wang, and Huazhu Fu. Cabnet: Category attention block for imbalanced diabetic retinopathy grading. *IEEE Transactions on Medical Imaging*, 2020.

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[23] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *arXiv preprint arXiv:1612.02649*, 2016.

[24] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, pages 4700–4708, 2017.

[25] Zilong Huang, Xinggang Wang, Jiasi Wang, Wenyu Liu, and Jingdong Wang. Weakly-supervised semantic segmentation network with deep seeded region growing. In *CVPR*, pages 7014–7023, 2018.

[26] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *AAAI*, volume 33, pages 590–597, 2019.

[27] Hengyuan Kang, Liming Xia, Fuhua Yan, Zhibin Wan, Feng Shi, Huan Yuan, Huiting Jiang, Dijia Wu, He Sui, Changqing Zhang, et al. Diagnosis of coronavirus disease 2019 (covid-19) with structured latent multi-view representation learning. *IEEE Transactions on Medical Imaging*, 2020.

[28] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised

image segmentation. In *ECCV*, pages 695–711. Springer, 2016.

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012.

[30] Jogendra Nath Kundu, Rahul Mysore Venkatesh, Naveen Venkat, Ambareesh Revanur, and R Venkatesh Babu. Class-incremental domain adaptation. In *ECCV*. Springer, 2020.

[31] Xiaomeng Li, Hao Chen, Xiaojuan Qi, Qi Dou, Chi-Wing Fu, and Pheng-Ann Heng. H-denseunet: hybrid densely connected unet for liver and tumor segmentation from ct volumes. *IEEE Transactions on Medical Imaging*, 37(12):2663–2674, 2018.

[32] Yanghao Li, Naiyan Wang, Jianping Shi, Xiaodi Hou, and Jiaying Liu. Adaptive batch normalization for practical domain adaptation. *Pattern Recognition*, 80:109–117, 2018.

[33] Jian Liang, Yunbo Wang, Dapeng Hu, Ran He, and Jiashi Feng. A balanced and uncertainty-aware approach for partial domain adaptation. In *ECCV*, 2020.

[34] Dongnan Liu, Donghao Zhang, Yang Song, Fan Zhang, Lauren O'Donnell, Heng Huang, Mei Chen, and Weidong Cai. Unsupervised instance segmentation in microscopy images via panoptic domain adaptation and task re-weighting. In *CVPR*, pages 4243–4252, 2020.

[35] Quande Liu, Cheng Chen, Jing Qin, Qi Dou, and Pheng-Ann Heng. Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space. In *CVPR*, pages 1013–1023, 2021.

[36] Yun Liu, Yu-Huan Wu, Yunfeng Ban, Huifang Wang, and Ming-Ming Cheng. Rethinking computer-aided tuberculosis diagnosis. In *CVPR*, pages 2646–2655, 2020.

[37] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015.

[38] Alexander Lundervold and Arvid Lundervold. An overview of deep learning in medical imaging focusing on mri. *Zeitschrift für Medizinische Physik*, 29(2):102–127, 2019.

[39] Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.

[40] Yujin Oh, Sangjoon Park, and Jong Chul Ye. Deep learning covid-19 features on cxr using limited training data sets. *IEEE Transactions on Medical Imaging*, 2020.

[41] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[42] Fei Pan, Inkyu Shin, Francois Rameau, Seokju Lee, and In So Kweon. Unsupervised intra-domain adaptation for semantic segmentation through self-supervision. In *CVPR*, pages 3764–3773, 2020.

[43] Pedro O Pinheiro and Ronan Collobert. From image-level to pixel-level labeling with convolutional networks. In *CVPR*, pages 1713–1721, 2015.

[44] Clément Playout, Renaud Duval, and Farida Cheriet. A novel weakly supervised multitask architecture for retinal lesions segmentation on fundus images. *IEEE Transactions on Medical Imaging*, 38(10):2434–2444, 2019.

[45] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[46] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[47] Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. Maximum classifier discrepancy for unsupervised domain adaptation. In *CVPR*, pages 3723–3732, 2018.

[48] Baochen Sun, Jiashi Feng, and Kate Saenko. Return of frustratingly easy domain adaptation. In *AAAI*, 2016.

[49] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018.

[50] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *CVPR*, pages 2097–2106, 2017.

[51] Botong Wu, Xinwei Sun, Lingjing Hu, and Yizhou Wang. Learning with unsure data for medical image diagnosis. In *ICCV*, 2019.

[52] Hongtao Xie, Dongbao Yang, Nannan Sun, Zhineng Chen, and Yongdong Zhang. Automated pulmonary nodule detection in ct images using deep convolutional neural networks. *Pattern Recognition*, 85:109–119, 2019.

[53] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *CVPR*, pages 2272–2281, 2017.

[54] Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 2020.

[55] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *CVPR*, pages 2921–2929, 2016.

[56] Yi Zhou, Xiaodong He, Lei Huang, Li Liu, Fan Zhu, Shanshan Cui, and Ling Shao. Collaborative learning of semi-supervised segmentation and classification for medical images. In *CVPR*, pages 2079–2088, 2019.

[57] Yi Zhou, Lei Huang, Tianfei Zhou, and Ling Shao. Many-to-one distribution learning and k-nearest neighbor smoothing for thoracic disease identification. In *AAAI*, volume 35, pages 768–776, 2021.

[58] Yi Zhou, Boyang Wang, Lei Huang, Shanshan Cui, and Ling Shao. A benchmark for studying diabetic retinopathy: Segmentation, grading, and transferability. *IEEE Transactions on Medical Imaging*, 40(3):818–828, 2020.

[59] Yi Zhou, Tianfei Zhou, Tao Zhou, Huazhu Fu, Jiacheng Liu, and Ling Shao. Contrast-attentive thoracic disease recognition with dual-weighting graph reasoning. *IEEE Transactions on Medical Imaging*, 40(4):1196–1206, 2021.

[60] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming 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*, pages 3–11. Springer, 2018.

[61] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, pages 2223–2232, 2017.