Debiased Curriculum Adaptation for Safe Transfer Learning in Chest X-ray Classification

Supplementary Material

A. Full Results

A.1. Main Results

Due to the limited pages, we list the experimental results for Consolidation in Tab. 9, the experimental results for Pleural Effusion in Tab. 10, and the experimental results for Edema in Tab. 11.

A.2. Additional Results on Dermoscopic Dataset

To further validate the potential of DCA's extension to other medical datasets, we conducted additional experiments on the dermatology dataset. HAM [42] and MSK [10] datasets from the widely used ISIC 2019 challenge are employed to verify performance on 3 diseases based on dermoscopic images. As in Tab 8, DCA achieves SOTA performance on three skin diseases, proving the potential to be extended to other types of medical imaging.

Table 8. From HAM to MSK for three diseases in ISIC challenge.

Disease	Melanoma	Dermatofibroma	Vascular lesion	Average
DANN	82.349	85.010	91.733	86.364
CDAN	81.747	84.785	94.669	87.067
MCD	82.898	82.339	92.096	85.778
MDD	84.127	89.278	92.859	88.755
DCA	**85.883**	**91.233**	**98.604**	**91.907**

B. Experimental Details

B.1. Dataset Processing

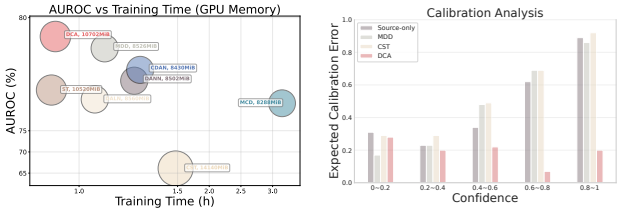
To maintain consistency across the four datasets, we filter out lateral chest X-ray images, retaining only frontal images. Each dataset is randomly divided into 80% training, and 20% testing. Since individual patients have multiple follow-up acquisitions, all data from a patient is assigned to a single subset only.

B.2. Implementation Details

We use **PyTorch** to implement our methods and fine-tune ResNet-50 pre-trained on ImageNet [37]. Following the standard protocols for unsupervised domain adaptation [34], all labeled source samples and unlabeled target samples participate in the training stage. We adopt mini-batch SGD with momentum of 0.9 and use the learning rate schedule of DANN [17], assigning a smaller learning rate $lr = 0.001$ to the feature extractor and a larger learning rate $lr = 0.01$ to the classifier. We resize the images to 256×256. Random cropping and horizontal flipping are used as data augmentation techniques. The evaluation criterion is Area Under the Receiver Operating Characteristic (AUROC) [16] curve score. γ in the overall optimization function (Equation 10 in the paper) trades off the importance of curriculum adaptation and spectral debiasing, which is set 0.0002 in all experiments. To select pseudo labels with high confidence, we need two fixed thresholds. In the case of imbalance, thresholds of different classes often differ significantly. In DCA, the proposed Class-wise Spectral Debiasing, can effectively alleviate the problem by balancing the feature space of positive and negative classes. Therefore, in all experiments, we uniformly set the threshold of negative class (with label $y = 0$) to 0.1 and the threshold of positive class (with label $y = 1$) to 0.9.

C. More Analysis

Computational Complexity The proposed Spectral Debiasing (SD) requires additional computational costs. SD computes the singular values on the feature matrix of a minibatch with the complexity of $O\left(b^2 d\right)$, where d is the dimension of features and b is the batch size. Since b is often small, the overall computational budget of computing singular values is nearly negligible in the mini-batch SGD training. As shown in Fig. 6a, DCA achieves optimal performance with sub-optimal training time cost and controllable memory overhead.



(a) Computational Cost (b) Confidence Calibration

Figure 6. (a) Computational cost comparison with baselines. (b) Confidence calibration: ECE of different confidence bins.

Calibration Analysis DCA incorporates implicit confidence calibration. We use Expected Calibration Error (ECE) [20] to investigate this issue on the transfer task from CheXpert to NIH-CXR for Atelectasis. Confidence is defined as: confidence = $2 \times |0.5 - \text{predication}|$. The smaller the ECE, the more reliable the model's predictions. It can be observed from Fig. 6b that, compared to the baselines, DCA significantly reduces ECE, especially in the high-confidence bins, mitigating overconfidence issues and ensuring more reliable predictions.

Table 9. AUROC (%) on 12 transfer tasks across four domains for Consolidation.

Method	C→M	C→N	C→O	M→C	M→N	M→O	N→C	N→ M	N→O	O→C	O→M	O→N	Average
Source-only	72.670	75.782	86.904	63.158	74.968	87.563	63.368	72.843	88.933	56.078	63.045	59.986	72.108
DANN	**73.674**	74.925	85.033	63.793	75.314	86.008	62.980	73.674	90.593	52.448	59.081	59.998	71.460
CDAN	73.125	73.605	84.743	63.763	74.578	85.744	63.657	73.125	86.298	53.176	53.899	56.153	70.156
MCD	72.418	74.862	85.191	62.590	73.384	89.354	59.808	68.740	90.066	53.130	61.522	63.717	71.232
MDD	73.249	75.928	83.847	63.396	75.145	92.306	63.378	72.833	**93.570**	55.783	59.896	63.584	72.743
DALN	72.792	74.846	85.296	63.348	75.033	86.008	63.655	71.971	88.906	54.091	58.314	58.102	71.030
ST	71.964	72.427	**90.777**	63.449	74.219	**96.153**	61.530	70.181	89.644	56.703	63.071	62.772	72.741
CST	65.833	68.318	87.563	64.024	75.619	86.087	59.883	65.454	87.668	53.934	59.780	61.210	69.614
DCA (ours)	73.505	**76.020**	90.198	**64.250**	**76.140**	92.806	**64.454**	**74.154**	92.279	**59.189**	**68.303**	**70.332**	**75.136**

Table 10. AUROC (%) on 12 transfer tasks across four domains for Pleural Effusion.

Method	C→M	C→N	C→O	M→C	M→N	M→O	N→C	N→ M	N→O	O→C	O→M	O→N	Average
Source-only	88.722	87.404	94.687	85.438	87.736	98.667	82.495	87.586	95.675	76.022	81.696	80.649	87.231
DANN	88.470	85.654	88.544	85.165	87.515	97.935	81.622	87.557	96.245	67.954	72.077	74.768	84.459
CDAN	88.744	85.486	88.218	84.916	**87.753**	97.198	81.889	87.397	96.126	69.380	78.213	78.682	85.334
MCD	88.709	85.026	95.274	83.533	86.505	97.833	65.626	78.614	**97.008**	73.743	81.231	79.965	84.422
MDD	88.941	86.747	90.376	**85.659**	87.101	94.819	82.318	87.221	96.037	74.413	79.133	81.962	86.227
DALN	88.054	86.106	93.262	85.027	87.519	98.102	81.123	87.117	95.900	70.275	77.898	78.321	85.725
ST	86.664	86.504	92.516	83.709	87.552	**98.107**	79.642	84.074	96.236	75.311	74.846	76.382	85.129
CST	82.631	71.579	92.944	85.314	87.232	97.392	**83.881**	85.245	95.543	69.071	72.961	**84.396**	84.016
DCA (ours)	**88.941**	**87.597**	**96.165**	85.585	87.737	97.577	82.752	**88.035**	95.539	**75.592**	**81.436**	82.189	**87.429**

Table 11. AUROC (%) on 12 transfer tasks across four domains for Edema.

Method	C→M	C→N	C→O	M→C	M→N	M→O	N→C	N→ M	N→O	O→C	O→M	O→N	Average
Source-only	86.560	88.763	96.803	80.245	87.118	95.429	65.264	73.164	96.565	69.243	73.685	82.941	82.982
DANN	85.724	86.601	95.271	**81.104**	87.863	97.701	**72.681**	79.031	**98.494**	58.593	67.868	77.649	82.382
CDAN	85.893	85.642	96.909	80.873	87.553	97.728	71.061	79.263	97.781	57.452	58.625	70.391	80.764
MCD	86.233	87.323	96.513	76.084	87.761	**97.992**	61.640	72.876	97.199	55.178	56.332	66.454	78.465
MDD	**86.558**	88.169	97.173	80.894	86.989	95.773	63.583	75.520	**97.913**	68.707	72.404	81.078	82.897
DALN	86.099	87.651	96.803	80.131	87.671	95.139	70.036	76.937	94.003	68.552	72.932	81.065	83.085
ST	85.533	86.777	95.799	79.352	86.967	94.954	69.822	77.838	96.037	68.154	74.338	79.432	82.917
CST	84.799	63.524	95.165	80.435	82.101	88.692	68.469	77.168	79.894	65.275	65.630	77.195	77.362
DCA (ours)	86.400	**88.513**	**98.943**	80.440	**88.060**	97.120	72.601	**80.135**	96.645	**70.400**	**76.070**	**84.340**	**84.972**