A MINIMALLY SUPERVISED APPROACH FOR MEDICAL IMAGE QUALITY ASSESSMENT IN DOMAIN SHIFT SETTINGS

Huijuan Yang[⋆†], *Aaron S. Coyner*[†], *Feri Guretno*[⋆†], *Ivan Ho Mien*[⋆†], *Chuan Sheng Foo*[⋆†]
J. Peter Campbell[†], *Susan Ostmo*[†], *Michael F. Chiang*[◇] *and Pavitra Krishnaswamy*[⋆†]

[⋆†]Institute for Infocomm Research, A*STAR, Singapore;
[†]Oregon Health & Science University, USA; [◇]National Eye Institute, National Institutes of Health

ABSTRACT

Accurate disease diagnosis requires objective assessment of clinical image quality. Automated image quality assessment (IQA) could enhance screening and diagnosis workflows. However, development of generalizable quality assessment tools requires large labeled clinical image datasets from different sites. Obtaining these datasets is often infeasible; and quality indicators may vary with acquisition settings due to domain shift. We introduce a minimally-supervised image quality assessment (MIQA) approach that can learn effectively with small datasets and limited labels in class-imbalanced domain shift scenarios. We formulate the IQA task as an anomaly detection problem, and use a small number of target domain images to identify a compact subset of source domain data for better representation of acceptable quality features. For this compact source domain dataset, we extract features with a pre-trained CNN, perform adaptive feature selection, and develop a one-class classifier to detect poor quality images. We evaluate our approach on two ophthalmology datasets, and show substantial AUC gains and improved cross-site generalizability over competitive baselines. Our approach has implications for improved image quality audit in many clinical settings.

***Index Terms*—** Medical image quality assessment, minimally supervised, class imbalance, data scarcity, domain shift

1. INTRODUCTION

Image quality heavily affects diagnostic accuracy and precision in clinical practice. Hence, there is interest in automated tools to objectively assess clinical image quality. Although image quality assessment (IQA) by itself would not directly identify medical image abnormalities, it nevertheless aids in indicating the reliability of resultant inferences [1]. However, developing automated clinical image quality assessment tools is challenging because clinical image datasets have noise and artifacts, commonly exhibit normal-abnormal class imbalance, and show significant domain shifts across multiple acquisition sites [2, 3]. Digital fundus photographs of the human eye are used to screen for, detect and grade diseases such as diabetic retinopathy (DR) and retinopathy of prematurity (ROP). In these use cases, automated IQA can identify low-quality images which require reacquisition, saving time and resources while improving clinical care.

Conventional IQA methods for digital fundus images employ generic parameters such as focus, clarity and illumination, and structural parameters such as eye vasculature and blood vessels in ROIs [4]. The heavy reliance on segmentation quality limits applicability of structure-based methods. In contrast to handcrafted features used in conventional methods, supervised deep neural network (DNN)-based methods extract multi-level features from the DNN and transfer knowledge learned from pretrained models to the target tasks, e.g., by fine-tuning the selected layers [5, 6, 7, 8]. The following are some examples. To mitigate the class imbalance issue, supervised methods differentiate multi-center images of "acceptable quality" from those of "possibly acceptable" and "not acceptable" qualities [6, 7], demonstrating good correlation between prediction scores and scores from expert graders. The capability to pinpoint locations of low quality has been investigated in patch-based methods by pooling the patch classification results via weighted average pooling [5]. Multiple color-space fusion network-based retinal IQA integrates the representations of different color-space at both feature-level and prediction-level [9]. Unsupervised image saliency map features (which may be combined with supervised features [10]) are employed to reflect how the human visual system identifies image quality to mitigate any observer's prejudiced judgment [11]. These supervised methods require large clinical image datasets comprising images of varying quality, acquired at different sites, all with associated quality-level annotations by experts. But obtaining these datasets is often infeasible.

Further, it is difficult to maintain performance of image quality assessment tools across acquisition settings. Often, data size, class balance, quality characteristics, and statistical distributions vary widely across acquisition devices and clinical sites. A prior study proposed semi-tied adversarial discriminative domain adaptation [12] for IQA, but this approach needs to detect the optic disc and fovea as landmarks to assist

coarse-to-fine feature encoding. Generally, it is known that conventional domain adaptation methods do not perform well for small medical dataset with significant data quality variations and distribution shift [13, 14].

To address the above challenges collectively, we focus on developing IQA tools that can learn effectively with small datasets and limited labels in domain shift scenarios.

1. We observe that the IQA task is often severely class-imbalanced as images with unacceptable quality tend to have very low prevalence in medical image datasets, and formulate clinical IQA as an anomaly detection problem.
2. We propose a practical, minimally-supervised image quality assessment (MIQA) method for anomaly detection under data and label scarcity scenarios with domain shift. To mitigate the domain shift across deployment sites, we use a small number of labeled target domain images to identify a compact subset of the source domain dataset with features of acceptable quality images; and then use this compact source domain dataset to train a one-class classifier for IQA.
3. We evaluate our method on two medical datasets, including a real-world multi-center ROP dataset with significant site-to-site quality variation, and show substantial performance gains over competitive baselines.

2. MINIMALLY-SUPERVISED IMAGE QUALITY ASSESSMENT (MIQA)

We consider a practical scenario where a development site has sizeable datasets and deployment sites generally have small datasets. The objective is to develop an automated tool for image quality assessment that will have robust performance at several deployment sites. We formulate this image quality assessment task as an anomaly detection problem. Ideally, we would address this task by training an anomaly detection model with a large number of good quality images from multiple sites. However, in reality, we may only have access to sizeable image datasets from one source site. Further, these source images may constitute a mix of different quality levels and have no quality labels. This combination of data scarcity, lack of labels, class imbalance and need to train generalizable models introduces significant challenges.

We evolve a minimally supervised approach to this problem as follows. We consider source (development) and target (deployment) domain datasets D_s and D_t. Due to the need to generalize models in the target domain, we sample a small number of images from D_t to label. We term the labelled subset of D_t as D_{tl}. We then use D_{tl} to identify a compact subset of D_s, $\hat{D}_s$, that represents features of good quality images. For this step, we use D_{tl} to train an ensemble of one-class classifiers and apply the resulting model on D_s to identify $\hat{D}_s$. This pruning results in a sizeable and clean development dataset comprising primarily good quality images. We then use $\hat{D}_s$ to develop an unsupervised classifier to detect deviations from acceptable diagnostic quality. Our overall method for minimally supervised image quality assessment (MIQA) is illustrated in Fig. 1.

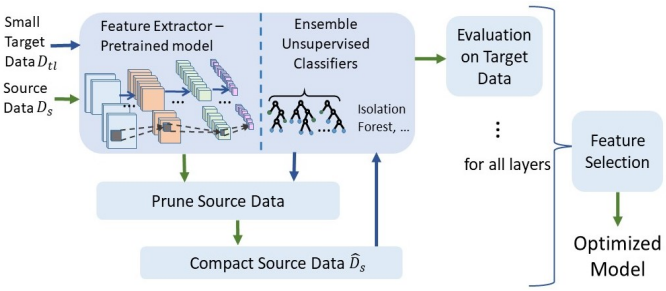


Fig. 1: Illustration of Proposed Minimally-Supervised Image Quality Assessment Method.

We now describe the algorithm for MIQA (pseudocode in Algorithm 1). First, we feed the small labelled target dataset (D_{tl}) to a pretrained CNN feature extractor parametrized by θ ($\mathcal{F}_e(\theta)$) and extract the features for layer l, i.e., f_{tl}^l. We then use f_{tl}^l to train an ensemble of one-class classifiers ($\mathcal{C}_k^l$, where k=1,2,...n_c). Next, we feed the source data D_s through $\mathcal{F}_e(\theta)$ to obtain features (f_s^l), and apply the trained ensemble of one-class classifiers on f_s^l to obtain anomaly scores for D_s. For each classifier in the ensemble, we threshold the anomaly scores and designate images with scores in the top p_s percentile (with allowance λ) as anomalous. Then, we average the predictions from the different classifiers, and adjudicate the top p_s percentile samples as having unacceptable diagnostic quality. We prune these samples and retain the remaining samples in D_s. We use this set of acceptable quality images, $\hat{D}_s$, to train our anomaly detection model. Finally, for each sample in $\hat{D}_s$, we extract features $\hat{f}_s^l$ for every layer $l = 1, \ldots, L_t$ of the pre-trained model. Then, each feature set gives rise to one unsupervised anomaly detection model. We evaluate performance for each feature set on the labelled target data D_{tl}, and select the one that provides the best performance as the final anomaly detection model for test-phase evaluation.

3. EXPERIMENTAL EVALUATION

3.1. Datasets

We evaluated our method on two medical fundus image quality datasets.

First, we used the public **Diabetic Retinopathy image Quality (DR-Quality)** dataset comprising 28,292 images taken from the original DR dataset [15] and re-annotated with image quality labels of "Good", "Usable", and "Reject" [9]. To assess the effectiveness of our method for classifying medical image quality with small imbalanced and unlabeled datasets, we sought to simulate scenarios of data and label scarcity, and class imbalance. Specifically, we generated 10 splits of varying sizes (i.e., with the no. of samples varying from 400 to 2500 images) via stratified random sampling

Algorithm 1 Pseudo Codes for Our Proposed Algorithm

Input: D_s: Source data with acceptable quality; D_{tl}, Y_{tl}: target validation data and label; $\mathcal{C}$, n_c: One-class classifiers, and total number of classifiers; P_t: predictor of outliers.
Parameter: p_s: outlier proportion; M_{pt}, $\mathcal{F}_e$: pretrained model with L_t layers, and feature extractor; λ: allowance factor
Output: Optimized unsupervised train model

for $l \in \{L_t$ layers of $M_{pt}\}$ **do**
  Compute features f_s^l and f_{tl}^l for D_s and D_{tl} by: $(f_s^l, f_{tl}^l) = \mathcal{F}_e(\theta, M_{pt}, L_t^l, D_s, D_{tl})$
  Train k one-class classifiers to obtain: $\mathcal{C}_k^l$.
  while for $k \in \{0,1,...,n_c\}$ **do**
    Train $\mathcal{C}_k^l$ by feeding f_{tl}^l
    Obtain outlier scores $S_c^k = \mathcal{C}_k^l(f_s^l, \lambda, p_s)$
    Predict labels of outlier: $O_s^k = P_t(S_c^k, p_s)$
  end while
  Compute adjudicated prediction of outlier by $\bar{O}_s = \frac{1}{n_s}\sum_{k=1}^{n_c} O_s^k$ and calculate threshold (t_s) for p_s percentile (C_p): $t_s = C_p(\bar{O}_s, p_s)$
  $I_s = \{i | v_i \in \bar{O}_s \wedge v_i > t_s\}$ and prune f_s^l: $\hat{f}_s^l = f_s^l[I_s]$.
  Train an one-class classifier $\mathcal{C}_m^l$ by feeding pruned features $\hat{f}_s^l$
  Predict f_{tl}^l using C_m^l to obtain performance P_{tl}^l
end for
Finally choose $\hat{l}$ by: $\hat{l} = \underset{l, l \in \{L_t\}}{\operatorname{argmax}}(P_{tl}^l)$
return trained model $C_m^{\hat{l}}$ for test data evaluation

from the train and test sets of the relabeled dataset. The distribution of both the source and target data follow the original data distribution with limited domain shift. We subsequently test the effectiveness of our method under such scenarios.

Second, we used a proprietary **Retinopathy Of Prematurity image Quality (ROP-Quality)** dataset, which was acquired through the multi-site NIH-funded Imaging and Informatics in ROP (i-ROP) study [7], approved by the respective Institutional Review Boards [7]. These images are associated with diagnostic quality labels of "Acceptable Quality (AQ)", "Possibly Acceptable Quality (PAQ)", or "Not Acceptable Quality (NAQ)" based on consensus rating by three independent annotators. We used posterior view images from any site with more than 300 images for our experiments, and had a total of 5 sites (denoted site 0,1,2,3,4 with 443, 609, 1305, 1475 and 1977 samples respectively). Fig. 2 shows examplar images from different acquisition sites, highlighting the substantial site-to-site variations in appearance of image quality. The ROP-Quality dataset exhibits substantial variability in distributions of patient demographics (race, ethnicity, birthweight) across sites. As eye images are heavily influenced by race and ethnicity, and birthweight is an indicator of disease and socioeconomic background, these demographic differences translate directly to differences in image features across sites. Further, there is variability in the skills of technicians acquiring the images across sites, and this correlates directly with the quality of images captured [7, 16]. We empirically quantified degree of domain shift by training a CNN classifier to predict whether a given image is from the source or target domain and computing the domain classification accuracy (related to H-divergence [17], a theoretical measure of domain shift). Across the 20 site pairs in the ROP-Quality dataset, the mean classification accuracy was 91.08±4.35%, attesting to significant domain shift amongst the different acquisition sites.

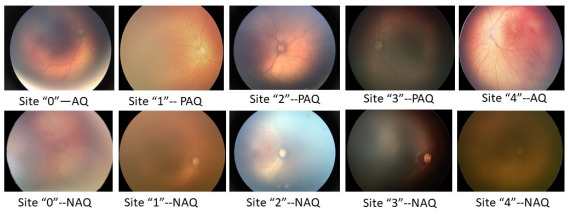


Fig. 2: Exemplar ROP Images from Different Acquisition Sites for ROP-Quality dataset.

For both the DR-Quality and ROP-Quality datasets, we designate the "Reject" (prevalence 18-21%) and "NAQ" (prevalence 1.7-11%) classes as the target anomalies for detection. For any designated target domain, we randomly sample 3% of the images (amounting to 12 to 75 images) while maintaining the class proportions for labeling.

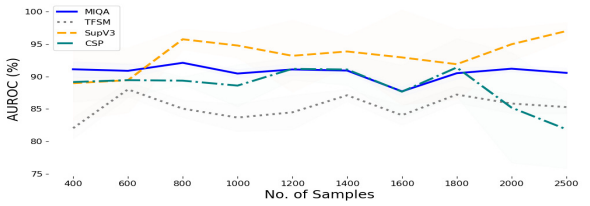
3.2. Benchmarking

We evaluate our proposed method against several baselines: Transfer Forest with feature selection based on Small validation data (TFSM), which is a high-performing unsupervised baseline [18]; two supervised methods based on Inception V3 model (SupV3) [7] and Color Space (CSP) [9] that have been previously applied for fundus IQA. Our method and Transfer Forest only require basic pixel normalization, while the supervised baselines require normalization and basic augmentation. We fine-tuned hyperparameters for SupV3 and CSP: learning rate: 0.01; batch size: 20 (SupV3-train), 6 (SupV3-test), 4 (CSP); epoch number: 20. For all baselines, we save the model with the highest validation performance, and evaluate AUROC and AUPRC on the designated target domain/test datasets. These metrics indicate the performance of discriminating between normal and anomalous classes. In particular, AUPRC measures the ability to detect the minority anomalous quality class given the heavy class imbalance in our data.

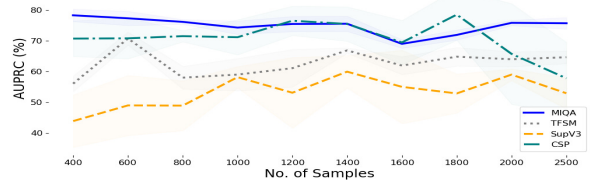
3.3. Results

The results for different splits and source-target pairings in the DR-Quality and ROP-Quality datasets are presented in Figure 3 and Figure 4, respectively. Table 1 summarizes the average performance of each method across the different splits and source-target pairings for DR-Quality and ROP-Quality. We observe that our proposed MIQA method adapts well to

the data scarcity, class imbalance, domain shift and variation in the data across sites. For the DR-Quality dataset, MIQA enables gains of 5.39% in AUROC over TFSM [18], 2.16% in AUROC over CSP [9], while being slightly inferior (2.62% lower) than SupV3 [7]. Importantly, MIQA boosts AUPRC more substantially, with gains of 12.27%, 21.74% and 4.19% in AUPRC over TFSM, SupV3 and CSP, respectively. These results demonstrate that MIQA is far more effective in detecting the anomalous or poor quality images in both datasets. For the ROP-Quality dataset, MIQA also offers substantial gains across the different source-target site pairings. On average, MIQA boosts AUROC by 1.44%, 4.24% and 9.17%; and AUPRC by 2.46%, 4.7% and 10% over TFSM, SupV3 and CSP, respectively.

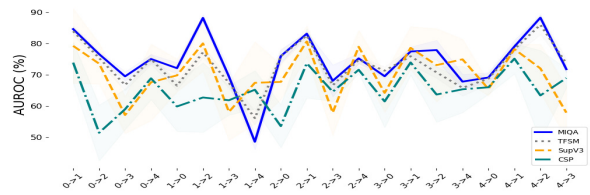


(a) DR-Quality: Comparison of AUROC.

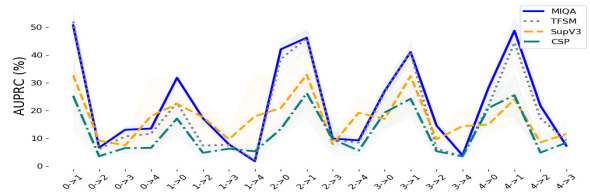


(b) DR-Quality: Comparison of AUPRC.

Fig. 3: Performance Comparison of Proposed Method With Baseline Methods for DR-Quality Data. Shaded areas represent standard deviations across seeds.



(a) ROP-Quality: Comparison of AUROC.



(b) ROP-Quality: Comparison of AUPRC.

Fig. 4: Performance Comparison of Proposed Method With Baseline Methods for ROP-Quality Data. Shaded areas represent standard deviations across seeds.

We note that MIQA performs well even in data scarce and class imbalance scenarios, whereas performance of the supervised methods depends on the size of source domain dataset. Further, we highlight that the performance of MIQA is very good even for the real-world multi-center ROP-Quality dataset which exhibits more serious domain shift, site-to-site variation, and class imbalance. This suggests its utility for practical settings that typically exhibit these challenges. Table 2 shows the best performing layers for the different site-to-site pairings in the ROP-Quality dataset. We see that MIQA typically adapts the selected layers (and hence feature representations) to the nature of the target data. The feature selection tends to prioritize lower visual layers when the target domain has more data (e.g., sites 3,4) and higher semantic layers when data in the target domains is more scarce.

Table 1: Average Performance Across Multiple Splits and Source-Target Pairs

DataSet	Method	Auroc (%)	Auprc (%)
DR-Quality	MIQA	90.66±1.14	**75.00±2.70**
	TFSm	85.27±1.83	62.73±4.43
	SupV3	**93.28±1.30**	53.26±5.08
	CSP	88.50±2.98	70.81±5.90
ROP-Quality	MIQA	**74.37±8.83**	**22.08±16.17**
	TFSm	72.93±7.15	19.62±16.10
	SupV3	70.13±8.10	17.38±8.23
	CSP	65.20±6.58	12.08±8.45

Table 2: Selected Layers ('SelLay') for Different Site Pairs: ROP-Quality Data

SelLay	Site Pairs	SelLay	Site pairs
Mixed_6a	2→0,2→3,1→2, 1→0,3→2	PrAuLogit	3→0, 4→0
	3→1,4→2,4→1	Conv2d_2a	1→4
MaxPool_5a_3x3	2→1,1→3, 0→1,3→4	Conv2d_2b	0→2
Mixed_5b	2→4,0→3,0→4	Conv2d_3b	4→3

4. CONCLUSIONS

In this paper, we presented MIQA, a Minimally-supervised Image Quality Assessment method for medical images. MIQA uses an anomaly detection framework to collectively address the challenges of data and label scarcity, class imbalance, and domain shift across acquisition sites for this problem. Specifically, we use a small target domain dataset to prune the source domain data and improve representation of features pertaining to acceptable quality, and train a one-class classifier to detect images of poor quality. In evaluations based on multi-center medical image quality datasets, we demonstrate that MIQA enables large performance gains over existing supervised and semi-supervised baselines. Our work has implications as a tool for automated image quality assessment in practical clinical AI deployments.

5. REFERENCES

[1] L.S. Chow and R. Paramesran, "Review of medical image quality assessment," *Biomedical Signal Processing and Control*, vol. 27, pp. 145–154, 2016.

[2] O. Dietrich, J. G. Raya, Scott B. Reeder, Maximilian F. Reiser, and Stefan O. Schoenberg, "Measurement of signal-to-noise ratios in mr images: Influence of multichannel coils, parallel imaging, and reconstruction filters," *Journal of Magnetic Resonance Imaging*, vol. 26, pp. 375–385, 2007.

[3] M. Fuderer, "The information content of mr images," *IEEE Trans. on Medical Imaging*, vol. 7, no. 4, pp. 368–380, 1988.

[4] J.M. Pires Dias, C.M. Oliveir, and L.A.d. Silva Cruz, "Retinal image quality assessment using generic image quality indicators," *Information Fusion*, vol. 19, no. 2014, pp. 73–90, 2014.

[5] P. Costa, A. Campilho, B. Hooi, A. Smailagic, K. Kitani, S. Liu, C. Faloutsos, and A. Galdran, "Eyequal: Accurate, explainable, retinal image quality assessment," in *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 2017, pp. 323–330, IEEE.

[6] A.S. Coyner, R. Swan, JM Brown, Jayashree Kalpathy-Cramer, Sang-Jin Kim, JP Campbell, K. Jonas, S. Ostmo, RVP Chan, and Michael F. Chiang, "Deep learning for image quality assessment of fundus images in retinopathy of prematurity," in *AMIA 2018 American Medical Informatics Association Annual Symposium, San Francisco, CA, Nov. 3-7*. 2018, AMIA.

[7] A.S. Coyner, R. Swan, J.P. Campbell, S. Ostmo, J.M. Brown, J. Kalpathy-Cramer, S.J. Kim, K.E. Jonas, RVP Chan, and M.F. Chiang, "Automated fundus image quality assessment in retinopathy of prematurity using deep convolutional neural networks," *Ophthalmology Retina*, vol. 3, no. 5, pp. 440–450, 2019.

[8] Jerone T. A. Andrews, Thomas Tanay, Edward J. Morton, and Lewis D. Griffin, "Transfer representation-learning for anomaly detection," *ICML Anomaly Detection Workshop*, 2016.

[9] H. Fu, B. Wang, J. Shen, S. Cui, Y. Xu, J. Liu, and L. Shao, "Evaluation of retinal image quality assessment networks in different color-spaces," in *Proc. of 22nd Int. Conf. on Medical Image Computing and Computer Assisted Intervention-MICCAI*. 2019, vol. 11764 of *Lecture Notes in Computer Science*, pp. 48–56, Springer.

[10] F. Yu, J. Sun, A. Li, J. Cheng, C. Wan, and J. Liu, "Image quality classification for DR screening using deep learning," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), Jeju Island, South Korea, July 11-15*. 2017, pp. 664–667, IEEE.

[11] Nabil G. Sadaka, Lina J. Karam, Rony Ferzli, and Glen P. Abousleman, "A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling.," in *ICIP*. 2008, pp. 369–372, IEEE.

[12] Y. Shen, B. Sheng, R. Fang, H. Li, L. Dai, S. Stolte, J. Qin, W. Jia, and D. Shen, "Domain-invariant interpretable fundus image quality assessment," *Medical Image Anal.*, vol. 61, pp. 101654, 2020.

[13] S. Lee, D. Kim, N. Kim, and S.-G. Jeong, "Drop to adapt: Learning discriminative features for unsupervised domain adaptation," in *IEEE/CVF Int. Conf. on Computer Vision, ICCV, Seoul, Korea*. 2019, pp. 91–100, IEEE.

[14] P. Morerio, J. Cavazza, and V. Murino, "Minimal-entropy correlation alignment for unsupervised deep domain adaptation," in *6th Int. Conf. on Learning Representations, ICLR, Vancouver, BC, Canada*. 2018, OpenReview.net.

[15] "Diabetic retinopathy detection," `https://www.kaggle.com/c/diabetic-retinopathy-detection/data`, 2015, [Online; accessed 27-September-2021].

[16] JS Chen, AS Coyner, S Ostmo, K Sonmez, S Bajimaya, E Pradhan, N Valikodath, ED Cole, T Al-Khaled, RVP Chan, P Singh, J Kalpathy-Cramer, MF Chiang, and JP Campbell, "Deep learning for the diagnosis of stage in retinopathy of prematurity: Accuracy and generalizability across populations and cameras," *Ophthalmol Retina.*, vol. 5, no. 10, pp. 1027–1035, 2021.

[17] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and VS Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, pp. 59:1–59:35, 2016.

[18] K. Ouardini, H. Yang, B. Unnikrishnan, M. Romain, C. Garcin, H. Zenati, JP Campbell, MF Chiang, J. Kalpathy-Cramer, V. Chandrasekhar, P. Krishnaswamy, and C.-S. Foo, "Towards practical unsupervised anomaly detection on retinal images," in *DART/MIL3ID@MICCAI*, 2019, vol. 11795, pp. 225–234.