Unimodal-uniform Constrained Wasserstein Training for Medical Diagnosis

Anonymous ICCV workshop (VRMI) submission

Paper ID 9

Abstract

The labels in medical diagnosis task are usually discrete and successively distributed. For example, the Diabetic Retinopathy Diagnosis (DR) involves five health risk levels: no DR (0), mild DR (1), moderate DR (2), severe DR (3) and proliferative DR (4). This labeling system is common for medical disease. Previous methods usually construct a multi-binary-classification task or propose some re-parameter schemes in the output unit. In this paper, we target on this task from the perspective of loss function. More specifically, the Wasserstein distance is utilized as an alternative, explicitly incorporating the inter-class correlations by pre-defining its ground metric. Then, the ground metric which serves as a linear, convex or concave increasing function w.r.t. the Euclidean distance in a line is explored from an optimization perspective.

Meanwhile, this paper also proposes of constructing the smoothed target labels that model the inlier and outlier noises by using a unimodal-uniform mixture distribution. Different from the one-hot setting, the smoothed label endues the computation of Wasserstein distance with more challenging features. With either one-hot or smoothed target label, this paper systematically concludes the practical closed-form solution. We evaluate our method on several medical diagnosis tasks (e.g., Diabetic Retinopathy and Ultrasound Breast dataset) and achieve state-of-the-art performance.

1. Introduction

In the realm of medical diagnosis, there are numerous prediction tasks in which the output labels demonstrate high discrete and successive features. The problem of the health risk level could be a very example. Although it can be a continuous variable, it is often discretized $i.e.$, at several intervals in practices. In parallel with previous statements, the Diabetic Retinopathy Diagnosis (DR) labels the disease level into five rankings: no DR (0), mild DR (1), moderate DR (2), severe DR (3) and proliferative DR (4) [18]. This kind of labeling system has been widely accepted by The Breast Imaging-Reporting and Data System (BI-RADS), liver (LIRADS), gynecology (GIRADS), colonography (CRADS), etc.

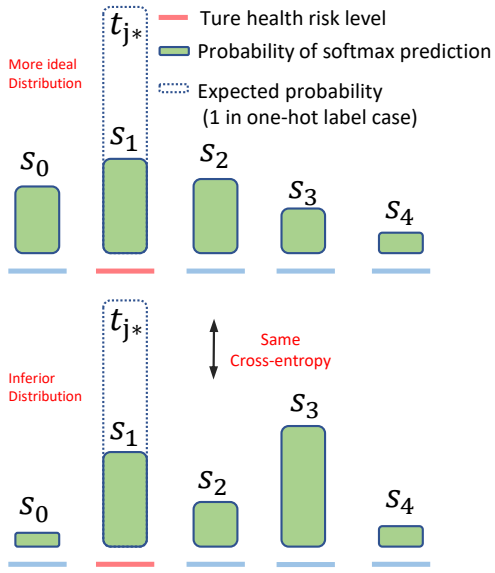

Figure 1. The limitation of CE loss for health risk level estimation. The ground truth direction of the car is t_j*. Two possible softmax predictions (green bar) of the health risk level estimator have the same probability at t_j* position. Therefore, both predicted distributions have the same CE loss. However, the top prediction is preferable to the bottom, since we desire the predicted probability distribution to be larger and closer to the ground truth class.

Due to the successively nature of risk level, the error that misclassifying a proliferative DR (4) image to mild DR (1) is considerably severe than the counterpart that misclassifying to severe DR (3). According to earlier literature, this kind of medical diagnosis task often casts as a multi-class classification problem or a metric regression problem.

The multi-class classification formulation using the cross-entropy (CE) loss, the class labels are assumed to be independent of each other [38]. Therefore, the inter-class similarity is not properly exploited. For instance, in Fig. 1, the predicted probability distribution is preferred to be concentrated near the ground truth class, while the CE loss does

not encourage that. On the other hand, metric regression methods treat the level as a continuous numerical value, although the label of health risk level itself is discrete. As manifested in [30, 11, 33, 28, 16], learning a regression model using discrete labels will cause over-fitting and exhibit similar or inferior performance compared with classification. Therefore, it is important to consider the successive and discrete nature of the health risk level classification problem.

Recent works typically choose the $N-1$ binary classification sub-tasks using sigmoid output with MSE loss or softmax output with CE loss, where N is the number of health risk levels. Unfortunately, the cumulative probabilities $p(y>1|\mathbf{x}),...,p(y>N-1|\mathbf{x})$ are calculated by several independent branches, failing to guarantee the that they are monotonically decreasing. This leads to the $p(y=i|\mathbf{x})$ are not ensured to be strictly positive, also the poor learning efficiency in the early stage of training. Moreover, $N-1$ weights need to be manually fine-tuned to balance the CE loss of each branch. [25] proposes to use a stick-breaking process to re-parametrize the outputs of $N-1$ units, which is closely related to the Bayesian non-parametric Dirichlet process. Embarking on this, the cumulative probabilities realize the expected monotonical decrease, but it could not be ignored that it is significantly more sophisticate than conventional CE-loss.

Furthermore, [2, 14, 3] propose to use a single output neuron to calculate the parameter of a unimodal distribution, and strictly require that the $p(y=i|\mathbf{x})$ follows a Poisson or Binomial distribution, which suffers from lacking the ability to control the variance [3]. Since the peak (also the mean and variance) of a Poisson distribution is equal to a designated λ, the peak cannot be assigned to the first or last class, and its variance is intended to be rather high when the peak is needed in the very later classes.

In this paper, we employ the Wasserstein loss as an alternative for empirical risk minimization. The 1^{st} Wasserstein distance is defined as the cost of optimal transport for moving the mass in one distribution to match the target distribution [6, 40, 41]. Specifically, the Wasserstein distance, between softmax prediction and its target label that are separately normalized as histograms, is measured. By defining the ground metric as class similarity, prediction performance earns the measure room in a sensitive way to correlations between the classes.

The ground metric can be predefined when the similarity structure is known as a priori to incorporate the inter-class correlation, $e.g.$, the Euclidean distance in a line. We further extend the Euclidean distance to its increasing function from an optimization perspective. The exact Wasserstein distance in a one-hot target label setting can be formulated as a soft-attention scheme of all prediction probabilities and be rapidly computed.

Another challenge of health risk level estimation stems from the label quality. For instance, the agreement rate of the radiologists for malignancy is usually less than 80%, resulting in a noisy labeled dataset [32, 42]. Despite the often-uncleared distinction between adjacent labels, it is more possible that a well-trained annotator will mislabel a Severe DR (3) sample to Moderate DR (2) rather than No DR (0). This requires modeling the noise for robust training [21].

The wrongly annotated targets may bias the training process [44, 4, 5, 35]. We instigate two types of noise. The outlier noise corresponds to one training sample being very distant from others by random error, and can be modeled by a uniform distribution [44]. It is noticed that the health risk level data is more likely to have inlier noise where the labels are wrong annotated as the near levels, and propose to model it by using a unimodal distribution. This paper proposes the solution which is to construct a smoothed target distribution by smoothing the one-hot label using a uniform-unimodal mixture model.

Unlike the one-hot setting, the smoothed target distribution makes the computation of Wasserstein distance more advanced because of the numerous possible transportation plans. The $\mathcal{O}(N^3)$ computational complexity for N classes has long been a stumbling block in using Wasserstein distance for large-scale applications. Instead of only obtaining its approximate solution using a $\mathcal{O}(N^2)$ complexity algorithm [13, 17], we systematically analyze the fast closed-form computation of Wasserstein distance for the smoothed label, in the context that the ground metric is a linear, convex, or concave increasing function $w.r.t.$ the Euclidean distance. The linear and convex cases can be solved with a linear complexity of $\mathcal{O}(N)$. In comparison to approximate counterpart, $exact$ solutions that this paper proposes are more effective.

The main contributions of this paper are summarized as follows:

• The health risk level estimation casts as a Wasserstein training problem. The inter-class relationship of health risk level data is explicitly incorporated as prior information in the gained ground metric which can be pre-defined ($e.g.$, a function $w.r.t.$ Euclidean distance in a line).

• The inlier and outlier error of health risk level is modeled using a discrete unimodal-uniform mixture distribution, and regularizes the target confidence by transforming one-hot label to the smoothed target label.

• For either one-hot or smoothed target label, this paper systematically concludes the possible fast closed-form solution when a non-negative linear, convex or concave increasing mapping function is applied in ground metric.

We empirically validate the effectiveness and generality of the proposed method on multiple challenging benchmarks. The Wasserstein loss obtaining superior performance over the current methods, especially using convex

mapping function for ground metric, smoothed label, and closed-form solution.

2. Related Works

2.1. Health risk level estimation

The conventional health risk level estimation approaches can be classified into three classes, $i.e.$, naive, binary decomposition and threshold methods [19, 27, 48]. Moreover, the health risk is not the only domain that has the successive label. The age prediction and atheistic rating also tightly related to this task. Following the development of deep learning, several works have been proposed to target the successive data. [25, 37] put forward the multi-task learning framework. However, the percentages of each class are not guaranteed to be positive, which may hurt the training, especially that in the early stage. Besides, there are $N-1$ weights to balance the branches, which is a hard task for manually tuning. [25] proposes a sophisticated stick-breaking process to reparameterise the $N-1$ outputs to alleviate this issue. [26] incorporate the metric learning for data relationship analysis. Different from these methods, we propose to use the Wasserstein distance as the optimization objective to inherit the label similarity.

2.2. Wasserstein distance

Wasserstein distance is a measure defined between probability distributions on a given metric space [46]. Recently, it attracted much attention in generative models etc [1]. [17] introduces it for multi-class multi-label task with a linear model. Given the significant amount of computing needed to solve the exact distance for general cases, these methods choose the approximate solution, of which the complexity is still in $\mathcal{O}(N^2)$ [13]. The fast computing of discrete Wasserstein distance is also closely related to SIFT [34, 10] descriptor, hue in HSV or LCH space [9, 15] and sequence data [29, 43]. Inspired by the above works, we further adapt this idea to the heath risk level estimation, and encode the geometry of label space by means of the ground matrix. We show that the fast algorithms exist in our pose label structure using the one-hot or smoothed target label and the ground metric is not limited to the Euclidean distance.

2.3. Robust training with noise data

Robust training with noise data has long been studied for general classification problems [21, 4]. Smoothing the one-hot label [44] with a uniform distribution or regularizing the entropy of softmax output [35] are two popular solutions. Some works of regression-based localization model the uncertainty of point position in a plane with a 2D Gaussian distribution [31, 45]. However, for the discrete successive label, the studies speak little voice.

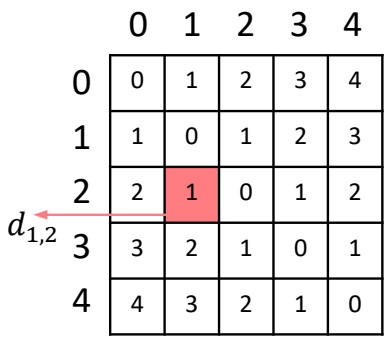

Figure 2. The ground matrix using Euclidean distance as ground metric.

2.4. Unimodality of Discrete and Successive Data

[14, 3] propose to enforcing the prediction to be a Poisson distribution. In their parametric version, the output of the neural network is a single sigmoid unit, which represents the parameter λ in Poisson distribution. However, requiring the output strictly follows a specific distribution could be a strong assumption. Besides, it is difficult to control the variance of the resulting Poisson distribution. [3] introduces an additional temperature parameter to control the variance, but results in more complicate hyper-parameter tuning. Here, we propose to use an exponential function following the softmax to flexibly adjust the shape of target label distribution. Noticing this modification could work on the target label rather than output distribution as [3].

3. Methodology

We consider the task of learning a health risk level estimator h_θ, parameterized by θ, with N-dimensional softmax output unit. It maps a medical image $\mathbf{x}$ to a vector $\mathbf{s} \in \mathbb{R}^N$. We perform learning over a hypothesis space $\mathcal{H}$ of h_θ. By virtue of the input $\mathbf{x}$ and its target ground truth one-hot label $\mathbf{t}$, typically, learning is performed via empirical risk minimization to solve $\min_{h_\theta \in \mathcal{H}} \mathcal{L}(h_\theta(\mathbf{x}), \mathbf{t})$, with a loss $\mathcal{L}(\cdot,\cdot)$ acting as a surrogate of performance measure.

Unfortunately, cross-entropy, information divergence, Hellinger distance and $\mathcal{X}^2$ distance-based loss treat the output dimensions independently [17], ignoring the similarity structure on pose label space.

Let $\mathbf{s} = \{s_i\}_{i=0}^{N-1}$ be the output of $h_\theta(\mathbf{x})$, $i.e.$, softmax prediction with N classes (health risk levels), and define $\mathbf{t} = \{t_j\}_{j=0}^{N-1}$ as the target label distribution, where $i, j \in \{0, \cdots, N-1\}$ be the index of dimension (class). Assume class label possesses a ground metric $\mathbf{D}_{i,j}$, which measures the semantic similarity between i-th and j-th dimensions of the output. There are N^2 possible $\mathbf{D}_{i,j}$ in a N class dataset and form a ground distance matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$. When $\mathbf{s}$ and $\mathbf{t}$ are both histograms, the discrete measure of exact Wasser-

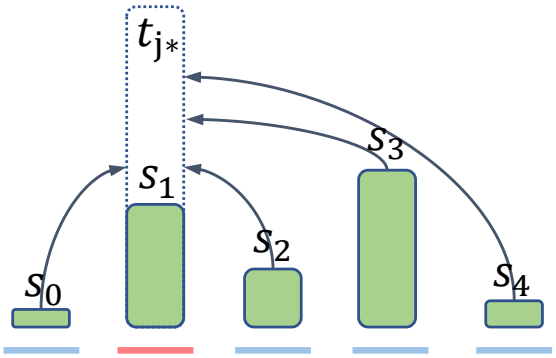

Figure 3. The only possible transport plan in one-hot target case.

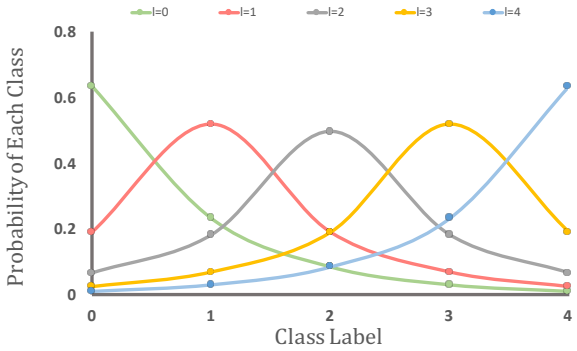

Figure 4. The distribution of normalized exponential function $e^{-|i-l|}$ for a dataset with 5 classes.

stein loss is defined as

$$\mathcal{L}_{\mathbf{D}_{i,j}}(\mathbf{s},\mathbf{t}) = \inf_{\mathbf{W}} \sum_{j=0}^{N-1}\sum_{i=0}^{N-1} \mathbf{D}_{i,j}\mathbf{W}_{i,j} \tag{1}$$

where $\mathbf{W}$ is the transportation matrix with $\mathbf{W}_{i,j}$ indicating the mass moved from the i^{th} point in source distribution to the j^{th} target position. A valid transportation matrix $\mathbf{W}$ satisfies: $\mathbf{W}_{i,j} \geq 0$; $\sum_{j=0}^{N-1}\mathbf{W}_{i,j} \leq s_i$; $\sum_{i=0}^{N-1}\mathbf{W}_{i,j} \leq t_j$; $\sum_{j=0}^{N-1}\sum_{i=0}^{N-1}\mathbf{W}_{i,j} = \min(\sum_{i=0}^{N-1} s_i, \sum_{j=0}^{N-1} t_j)$.

The ground distance matrix $\mathbf{D}$ in Wasserstein distance is usually unknown, but it has clear meanings in our application. Its i,j-th entry $\mathbf{D}_{i,j}$ could be the geometrical distance between the i-th and j-th points in a line. A possible choice is using the Euclidean distance $d_{i,j}$ of a line ($i.e.$, ℓ_1 distance between the i-th and j-th points in a line) as the ground metric $\mathbf{D}_{i,j} = d_{i,j}$.

$$d_{i,j} = |i-j| \tag{2}$$

The Wasserstein distance is identical to the Earth mover's distance when the two distributions have the same total masses ($i.e.$, $\sum_{i=0}^{N-1} s_i = \sum_{j=0}^{N-1} t_j$) and using the symmetric distance $d_{i,j}$ as $\mathbf{D}_{i,j}$. The ground matrix using Euclidean distance is shown in Fig.2.

This setting is satisfactory for comparing the similarity of SIFT or hue [10, 34, 40, 7, 47], which do not involve the neural network optimization. The previous efficient algorithm usually holds only for $\mathbf{D}_{i,j} = d_{i,j}$. Regarding this, this paper proposes extending the ground metric in $\mathbf{D}_{i,j}$ as $f(d_{i,j})$, where f is a positive increasing function $w.r.t.$ $d_{i,j}$.

3.1. Wasserstein training with one-hot target

For multi-class one-label dataset, the one-hot encoding is a typical setting. The distribution of a target label probability is $\mathbf{t} = \delta_{j,j^*}$, where j^* is the ground truth class, δ_{j,j^*} is a Dirac delta, which equals to 1 for $j = j^*$[1], and 0 otherwise.

[1]We use i,j interlaced for $\mathbf{s}$ and $\mathbf{t}$, since they index the same group of positions in a line.

Theorem 1. *Assume that* $\sum_{j=0}^{N-1} t_j = \sum_{i=0}^{N-1} s_i$, *and* $\mathbf{t}$ *is a one-hot distribution with* $t_{j^*} = 1$(or $\sum_{i=0}^{N-1} s_i$)[2], *there is only one feasible optimal transport plan.*

In keeping with the criteria of $\mathbf{W}$, all masses have to be transferred to the cluster of the ground truth label j^*, as illustrated in Fig. 3. Then, the Wasserstein distance between softmax prediction $\mathbf{s}$ and one-hot target $\mathbf{t}$ degenerates to

$$\mathcal{L}_{\mathbf{D}^f_{i,j}}(\mathbf{s},\mathbf{t}) = \sum_{i=0}^{N-1} s_i f(d_{i,j^*}) \tag{3}$$

where $\mathbf{D}^f_{i,j} = f(d_{i,j})$. f can be an increasing function proper, $e.g.$, p^{th} power of $d_{i,j}$ and Huber function. The exact solution of Eq. (3) can be computed with a complexity of $\mathcal{O}(N)$. The ground metric term $f(d_{i,j^*})$ works as the weights $w.r.t.$ s_i, which takes all classes into account following a soft attention scheme [24]. This explicitly encourages the probabilities distributing on the neighboring classes of j^*. Since each s_i is a function of the network parameters, differentiating $\mathcal{L}_{\mathbf{D}^f_{i,j}}$ $w.r.t.$ network parameters yields $\sum_{i=0}^{N-1} s_i' f(d_{i,j^*})$.

In contrast, the cross-entropy loss in one-hot setting can be formulated as $-1\log s_{j^*}$, which only considers a single class prediction like the hard attention scheme [24], which usually loses too much information. Similarly, the regression loss using softmax prediction could be $f(d_{i^*,j^*})$, where i^* is the class with maximum prediction probability.

3.2. Unimodal-uniform label smoothing

The outlier noise exists in most of the data-driven tasks, and can be modeled by a uniform distribution [44]. However, health risk labels are more possible to be mislabeled as a close class of the true class. It is more reasonable to construct a unimodal distribution to depict the inlier noise in

[2]We note that softmax cannot strictly guarantee the sum of its outputs to be 1 considering the rounding operation. However, the difference of setting t_{j^*} to 1 or $\sum_{i=0}^{N-1} s_i$) is not significant in our experiments using the typical format of softmax output which is accurate to 8 decimal places.

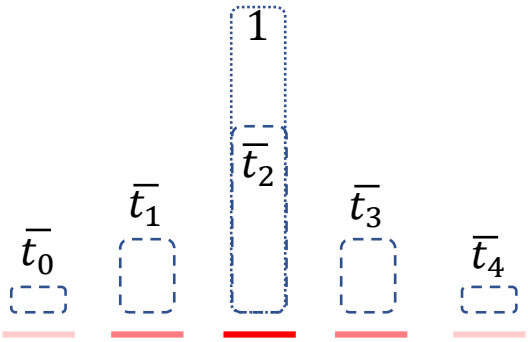

Figure 5. The unimodal-uniform smoothed target label distribution when the ground-of-truth class is 2.

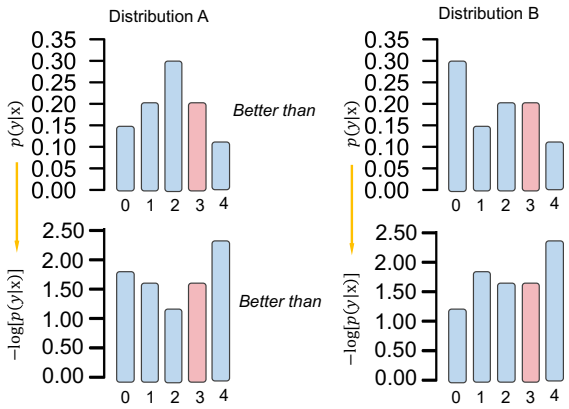

Figure 6. The expected (left) and inferior (right) distribution of predictions.

pose estimation, which has a peak at class j^* while decreasing its value for farther classes. This paper has sampled a continuous unimodal distribution and follow by normalization.

In here, we propose to sample on an exponential function $e^{\frac{-|i-l|}{\tau}}$ and followed by a softmax normalization. Discrete distributions with five classes are illustrated in Fig. 4.

The normalized unimodal value is denoted as p_j. The unimodal-uniform smoothed target distribution $\bar{\mathbf{t}}$ is constructed by replacing t_j in $\mathbf{t}$ with $(1-\xi-\eta)t_j+\xi p_j+\eta\frac{1}{N}$, which can be regarded as the weighted sum of the original label distribution $\mathbf{t}$ and a unimodal-uniform mixture distribution. In the context that the uniform distribution target is utilized for the CE loss, it is equivalent to label smoothing [44], a typical mechanism for outlier noisy label training, which encourages the model to accommodate less-confident labels. The smoothed distribution is shown in Fig. 5

It is noted that the smoothed target label can also be adapted into CE loss which is formulated as

$$\mathcal{L}=\sum_{i}^{N-1}\bar{t}_i[-\log(p(y=i|\mathbf{x}))] \tag{4}$$

By enforcing $\mathbf{s}$ to form a unimodal-uniform mixture distribution, we also implicitly encourage the probabilities to distribute on the neighbor classes of j^*.

Since $\bar{t}_i$ is monotonically decreasing $w.r.t.$ the farther distance from the true class l, we can regard it as a weight of $-\log(p(y=i|\mathbf{x}))$. From Fig. 6, this weight does consider the relative similarity of successive data, and has a smaller loss when the prediction probabilities are closer distribute around the l. Since the target label regularization can be processed in advance, the training time does not increase by adding the unimodal-uniform mixture distribution regularization.

3.3. Wasserstein training with smoothed target

With the unimodal-uniform smoothed target label, the fast computation of Wasserstein distance in Eq. (3) does not apply. A straightforward solution is to regard it as a general case and solve its closed-form result in the complexity higher than $\mathcal{O}(N^3)$ or get an approximate result with a complexity in $\mathcal{O}(N^2)$. The main results of this section are a series of analytic formulation when the ground metric is a nonnegative increasing linear/convex/concave function $w.r.t.$ Euclidean distance with reasonable complexity.

3.3.1 Linear and Convex function $w.r.t.$ $d_{i,j}$ as the ground metric.

When we use $d_{i,j}$ as ground metric directly or extend the ground metric as an nonnegative increasing and convex function of $d_{i,j}$, the analytic formulation of Wasserstein loss $\mathcal{L}_{d_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$ can be written as

$$\mathcal{L}_{d_{i,j}}(\mathbf{s},\bar{\mathbf{t}})=\sum_{j=0}^{N-1}|\sum_{i=0}^{j}(s_i-\bar{t}_i)| \tag{5}$$

To the best of our knowledge, Eq. (5) was first developed in [47], in which it is proved for sets of points with unitary masses ona line. A similar conclusion for the Kantorovich-Rubinstein problem was derived in [7, 8], which is known to be identical to the Wasserstein distance problem when $\mathbf{D}_{i,j}$ is a distance. We note that this is true for $\mathcal{L}_{d_{i,j}}$ (but false for $\mathcal{L}_{\mathbf{D}^\rho}(\mathbf{s},\bar{\mathbf{t}})$ with $\rho>1$). An equivalent distance is proposed from the cumulative distribution perspective [36]. All of these papers notice that computing Eq. (5) can be done in linear time ($i.e.$, $\mathcal{O}(N)$) weighted median algorithm (see [46] for a review).

We note that the partial derivative of Eq. (5) $w.r.t.$ s_n is $\sum_{j=0}^{N-1}\mathrm{sgn}(\varphi_j)\sum_{i=0}^{j}(\delta_{i,n}-s_i)$, where $\varphi_j=\sum_{i=0}^{j}(s_i-\bar{t}_i)$, and $\delta_{i,n}=1$ when $i=n$.

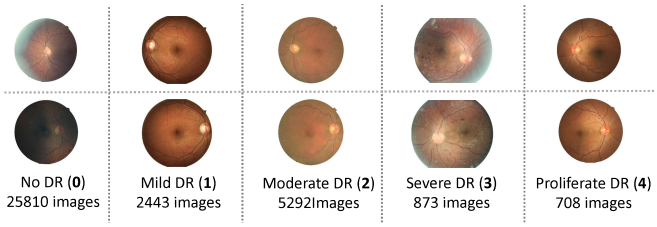

Figure 7. Some samples with different retinopathy level in the DR dataset. The top row is the left retinopathy image while the bottom row is the right retinopathy image. The sampls show a large inner-class variation and small inter-class variation.

Here, we give some measures [3] using the typical convex ground metric function.

$\mathcal{L}_{\mathbf{D}^{\rho}_{i,j}}(\mathbf{s}, \bar{\mathbf{t}})$, the Wasserstein measure using d^{ρ} as ground metric with $\rho = 2, 3, \cdots$. The case $\rho = 2$ is equivalent to the Cramér distance [39]. Note that the Cramér distance is not a distance metric proper. However, its square root is.

$$\mathbf{D}^{\rho}_{i,j} = d^{\rho}_{i,j} \tag{6}$$

$\mathcal{L}_{\mathbf{D}^{H\tau}_{i,j}}(\mathbf{s}, \bar{\mathbf{t}})$, the Wasserstein measure using a Huber cost function with a parameter τ.

$$\mathbf{D}^{H\tau}_{i,j} = \begin{cases} d^2_{i,j} & \text{if } d_{i,j} \leq \tau \\ \tau(2d_{i,j} - \tau) & \text{otherwise.} \end{cases} \tag{7}$$

3.3.3 Concave function $w.r.t.\ d_{i,j}$ as the ground metric

In practice, it may be useful to choose the ground metric as a nonnegative, concave and increasing function $w.r.t.\ d_{i,j}$. Although the general computation speed of the concave function is not satisfactory, the step function $f(t) = \mathbb{1}_{t \neq 0}$ (one every where except at 0) can be a special case, which has significantly less complexity [46]. Assuming that the $f(t) = \mathbb{1}_{t \neq 0}$, the Wasserstein metric between two normalized discrete histograms on N bins is simplified to the ℓ_1 distance.

$$\mathcal{L}_{\mathbb{1}_{d_{i,j} \neq 0}}(\mathbf{s}, \bar{\mathbf{t}}) = \frac{1}{2} \sum_{i=0}^{N-1} |\mathrm{s}_i - \bar{\mathrm{t}}_i| = \frac{1}{2} ||\mathbf{s} - \bar{\mathbf{t}}||_1 \tag{8}$$

where $|| \cdot ||_1$ is the discrete ℓ_1 norm.

Unfortunately, its fast computation is at the cost of losing the ability to discriminate the difference of probability in a different position of bins.

4. Experiments

To evaluate the effectiveness of our Wasserstein loss, we show implementation details and experimental results on the two widely used health risk level diagnosis datasets, $i.e.$, Diabetic Retinopathy and Ultrasound BIRADS datasets. To manifest the effectiveness of each setting choice and their combinations, we give a serial of elaborate ablation studies along with the standard measures. For the fair comparison, we choose the same neural network backbones as in previous works. All of networks in our training use the $\mathcal{L}_2$ norm of 10^{-4}, ADAM optimizer [22] with 128 training batch-size and initial learning rate of 10^{-3}. The learning rate will be divided by ten when either the validation loss or the valid set QWK plateaus. The is no significant difference in the training time of Wasserstein loss and CE-loss based multi-class classification, and the smoothed unimodal target label is constructed before the training stage. All the experiments are implemented in deep learning platform Pytorch [4].

We use the prefix $\approx$ denote the approximate computation of Wasserstein distance [13, 17]. $(\mathbf{s}, \mathbf{t})$ and $(\mathbf{s}, \bar{\mathbf{t}})$ refer to using one-hot or smoothed target label. For instance, $\mathcal{L}_{d_{i,j}}(\mathbf{s}, \mathbf{t})$ means choosing Wasserstein loss with Euclidean distance in a line as ground metric and using one-hot target label.

4.1. Evaluations

Since the health risk level has a discrete label, the performance of a system can be simply measured by the average classification accuracy as the conventional classification problem. [37] further utilized the Mean True Negative Rate (TNR) at True Positive Rate (TPR) of 0.95. The relatively high TPR used here is fitted for strict TPR requirements of medical applications to avoid misdiagnosing diseased cases as healthy. However, they do not consider the severity of different misclassification.

Considering the inherent ordered label relationship, the Mean Absolute Error (MAE) metric, $i.e.$, L_1 loss, can also be used as an evaluation metric in related risk evaluation datasets [33], which is computed using the average of the absolute errors between the ground truth and the estimated result. Here, we also propose its use in evaluating the proposed method on two medical health risk evaluation benchmarks.

Besides, following the previous metrics in the Kaggle competition of DR dataset, we choose the quadratic weighted kappa (QWK) [5] to implicitly punish the misclassification proportional to the distance between the ground-of-truth label and predicted label of the network [12]. The QWK is formulated as:

$$k = 1 - \frac{\sum_{i,j} \mathbf{W}_{i,j} \mathbf{O}_{i,j}}{\sum_{i,j} \mathbf{W}_{i,j} \mathbf{E}_{i,j}} \tag{9}$$

[3] We refer to "measure", since a ρ^{th}-root normalization is required to get a distance [46], which satisfies three properties: positive definiteness, symmetry and triangle inequality.

[4] https://pytorch.org/

[5] https://www.kaggle.com/c/diabetic-retinopathy-detection/overview/evaluation

Table 1. Performance on the DR dataset.

Evaluations	Mean TNR@TPR=0.95			Valid Acc	Valid QWK	MAE
	0 *vs* 1-4	0-1 *vs* 2-4	0-2 *vs* 3-4			
MC	41.5%	30.9%	31.1%	82.4%	0.724	0.37
RG	40.3%	30.6%	30.8 %	76.2%	0.705	0.38
Poisson [3]	38.8%	30.0%	29.6 %	77.1%	0.713	0.38
MT [37]	42.7%	31.7%	31.3%	82.8%	0.726	0.36
SB[25]	44.0%	33.1%	32.6%	84.2%	0.743	0.32
$\mathcal{L}_{d_{i,j}}(\mathbf{s},\mathbf{t})$	46.9%	37.1%	34.4%	87.3%	0.768	0.29
$\mathcal{L}_{\mathbf{D}^2_{i,j}}(\mathbf{s},\mathbf{t})$	47.2%	37.3%	34.6%	87.4%	0.769	0.28
$\mathcal{L}_{\mathbf{D}^{H_\tau}_{i,j}}(\mathbf{s},\mathbf{t})$	47.2%	37.4%	34.5%	87.6%	0.769	0.28
MC$(\mathbf{s},\bar{\mathbf{t}})$	42.4%	31.2%	31.8%	82.7%	0.728	0.35
$\approx \mathcal{L}_{d_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	45.8%	36.4%	33.8%	86.6%	0.759	0.29
$\approx \mathcal{L}_{\mathbf{D}^2_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	45.8%	36.5%	33.9%	86.5%	0.760	0.29
$\approx \mathcal{L}_{\mathbf{D}^{H_\tau}_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	45.9%	36.6%	34.0%	86.6%	0.760	0.28
$\mathcal{L}_{d_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	47.3%	37.5%	34.8%	87.8%	0.771	0.27
$\mathcal{L}_{\mathbf{D}^2_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	**47.6%**	**37.7%**	**34.9%**	87.8%	0.772	**0.26**
$\mathcal{L}_{\mathbf{D}^{H_\tau}_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	47.5%	**37.7%**	34.8%	**88.0%**	**0.773**	**0.26**

to measures the level of disagreement between two raters ($\mathcal{A}$ and $\mathcal{B}$). In here, the $\mathcal{A}$ is the $argmax$ prediction of our classifier and $\mathcal{B}$ is the ground truth. The $\mathbf{W}$ is a $N \times N$ matrix where $\mathbf{W}_{i,j}$ denotes the cost associated with misclassifying label i as label j. In QWK, $\mathbf{W}_{i,j} = (i-j)^2$. $\mathbf{O}_{i,j}$ counts the number of images that received a rating i by $\mathcal{A}$ and a rating j by $\mathcal{B}$. The quadratic calculation is one possible choice and one can plug in other distance metrics into kappa calculation. The matrix of expected ratings $\mathbf{E}$, is calculated, assuming that there is no correlation between rating scores. As a result, k is a scalar in [-1,1], and $k = 1$ indicates the two raters are total agreement, whereas $k < 0$ means the classifier performs worse than random choice. We make use of two typical health risk evaluation datasets in the medical area suitable for DNN implementations.

4.2. Diabetic Retinopathy (DR)

The Diabetic Retinopathy (DR) dataset [6] contains a large amount of high-resolution fundus ($i.e.$, interior surface at the back of the eye) images which have been labeled as five levels of DR. The level 0 to 4 representing the No DR, Mild DR, Moderate DR, Severe DR, and Proliferative DR, respectively. The left and right fundus images from 17563 patients are publicly available. The ResNet [20] style model with 11 ResBlocks as in [3, 25] has been adopted for DR dataset. We use five neurons with softmax normalization as our output to represent the probability of each level.

In our experiments, we follow the setting of [3, 25]. The subject-independent 10-fold cross-validation is adopted, $i.e.$, the validation set consisting of 10% of the patients is set aside. The images belonging to a subject will only appear in a single fold. By doing this, we can avoid contamination. The images are also preprocessed as in [2, 3, 25] and subsequently resized as 256×256 size images. Some examples can be found in Fig. 7.

We show the results in the DR dataset in Table 1. The evaluation metrics discussed earlier is utilized. Several baseline methods are chosen for comparisons. For example, the CE-loss based multi-class classification (MC), MSE-loss based metric regression (MSE), Poisson distribution output with CE-loss (Poisson), multi-task network using a series of binary CE loss (MT), and the stick-breaking with CE-loss (SB).

The MC usually outperforms MSE in most of the metric. However, MSE usually appears to be competitive w.r.t. MAE, since MSE optimizes a similar metric as MAE in its training phase. The Poisson does not manage to achieve performance improvements in most of the evaluations due to its uncontrollable variance. The MT is more promising than MC as it considers the successive relationship, despite it has a lot to be tuned hyper-parameters. By addressing some limitation in MT, the SB has a better performance than MT.

Our Wasserstein training outperforms all of the previous methods, especially the $\mathcal{L}_{\mathbf{D}^2_{i,j}}$ and $\mathcal{L}_{\mathbf{D}^{H_\tau}_{i,j}}$ which use the convex function of the Euclidean distance in a line as the ground metric.

Moreover, the unimodal-uniform smoothed target label can efficiently improve the performance without additional training costs. The smoothing process is benefiting to both conventional CE loss and Wasserstein loss. Besides, the exact solution can outperform its approximate counterpart

[6] https://www.kaggle.com/c/diabetic-retinopathy-detection

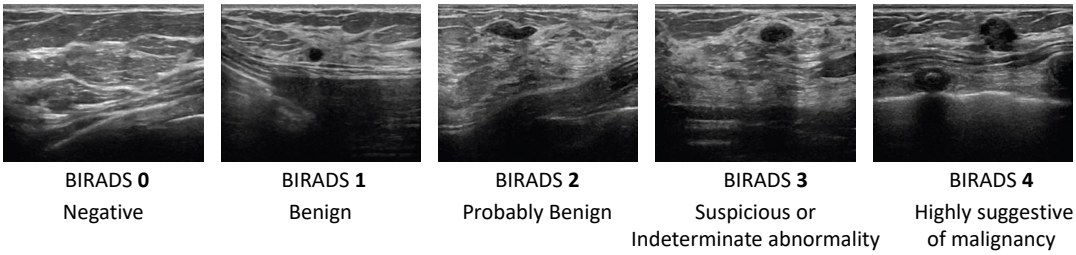

Figure 8. Some samples with different malignant risk in the US-BIRADS.

Table 2. Performance on the US-BIRADS dataset.*Our implementations have slightly higher TNR using MC baseline than the results reported in [37]

.

Evaluations	Mean TNR@TPR=0.95			Valid Acc	Valid QWK	MAE
	0 *vs* 1-4	0-1 *vs* 2-4	0-2 *vs* 3-4			
MC	33.2%*	28.7%*	29.8%*	73.3%	0.678	0.42
RG	31.6%	28.5%	29.5%	73.0%	0.677	0.44
Poisson [3]	29.6%	27.2%	29.5%	72.2%	0.665	0.45
MT [37]	38.5%	29.2%	31.3%	76.5%	0.685	0.41
SB[25]	39.1%	30.2%	32.0%	78.3%	0.694	0.39
$\mathcal{L}_{d_{i,j}}(\mathbf{s},\mathbf{t})$	42.5%	33.6%	35.7%	80.1%	0.712	0.36
$\mathcal{L}_{\mathbf{D}^2_{i,j}}(\mathbf{s},\mathbf{t})$	42.6%	33.8%	35.9%	80.2%	0.714	0.35
$\mathcal{L}_{\mathbf{D}^{H\tau}_{i,j}}(\mathbf{s},\mathbf{t})$	42.6%	33.7%	35.9%	80.3%	0.715	0.35
MC$(\mathbf{s},\bar{\mathbf{t}})$	33.4%	29.0%	30.4%	73.6%	0.682	0.40
$\mathcal{L}_{d_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	42.9%	34.0%	36.2%	80.5%	0.715	0.34
$\mathcal{L}_{\mathbf{D}^2_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	**43.0%**	**34.2%**	**36.3%**	80.5%	**0.716**	0.34
$\mathcal{L}_{\mathbf{D}^{H\tau}_{i,j}}(\mathbf{s},\bar{\mathbf{t}})$	**43.0%**	34.1%	**36.3%**	**80.6%**	**0.716**	**0.33**

consistently.

We set our hyper-parameters $\xi = 0.15$, $\eta = 0.05$ and $\tau = 1$. QWK is not sensitive to the $\tau \in \{0.8, 0.9, 1, 1.1\}$ when we fix the $\xi = 0.15$. Similarly, the QWK keep at the same level when we adjust ξ from 0.12 to 0.18.

4.3. Ultrasound BIRADS

The second medical dataset is the Ultrasound BIRADS (US-BIRADS) [37]. It consists of 4904 breast images with the BIRADS system label. Considering the relatively limited number of samples in level 4, we usually regard the 3-4 as a single level [37]. That results in 2700 healthy (0) images, 1113 benign (1) images, 359 probably benign (2) images, and 732 may contain/contain malignant (3-4) images. We divide this dataset into 5 subsets for subject-independent five-fold cross-validation. We show some samples at different levels in Fig. 8.

AlexNet style architecture [23] with six convolution layers and following two dense layers is used for US-BIRADS image dataset as in [37]. We set $\xi = 0.15$, $\eta = 0.05$, and $\tau = 1$ for the unimodal distribution.

The leading performance of our method is also observed in the US-BIRADS dataset (Table 2). Since its labels are noisier (more severe annotator-dependent problem), the unimodal-uniform smoothing usually offers a more appealing contribution to the results. The Wasserstein training with convex ground metric function and smoothed target label achieves state-of-the-art performance consistently.

5. Conclusions

Based on the Wasserstein distances, this paper has introduced a simple yet efficient loss function for health risk level estimation. Its ground metric represents the class correlation and itself can be predefined by using an increasing function of the Euclidean distance of a line. Both the outlier and inlier noise in pose data are incorporated in a unimodal-uniform mixture distribution to construct the conservative label. This paper systematically discusses the fast closed-form solutions in one-hot and conservative label cases. The results show that the best performance can be achieved by choosing convex function, unimodal-uniform distribution for smoothing and solving its exact solution. Although it was originally developed for health risk level estimation, it is essentially applicable to other problems with discrete and ordinal labels. In the future, we intend to develop an adaptive ground metric learning scheme, and adjust the shape of conservative target distribution automatically.

References

[1] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein gan. *arXiv preprint arXiv:1701.07875*, 2017. 3

[2] C. Beckham and C. Pal. A simple squared-error reformulation for ordinal classification. *arXiv preprint arXiv:1612.00775*, 2016. 2, 7

[3] C. Beckham and C. Pal. Unimodal probability distributions for deep ordinal classification. *arXiv preprint arXiv:1705.05278*, 2017. 2, 3, 7, 8

[4] A. J. Bekker and J. Goldberger. Training deep neural-networks based on unreliable labels. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 2682–2686. IEEE, 2016. 2, 3

[5] V. Belagiannis, C. Rupprecht, G. Carneiro, and N. Navab. Robust optimization for deep regression. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2830–2838, 2015. 2

[6] V. I. Bogachev and A. V. Kolesnikov. The monge-kantorovich problem: achievements, connections, and perspectives. *Russian Mathematical Surveys*, 67(5):785–890, 2012. 2

[7] C. A. Cabrelli and U. M. Molter. The kantorovich metric for probability measures on the circle. *Journal of Computational and Applied Mathematics*, 57(3):345–361, 1995. 4, 5

[8] C. A. Cabrelli and U. M. Molter. A linear time algorithm for a matching problem on the circle. *Information processing letters*, 66(3):161–164, 1998. 5

[9] S.-H. Cha. A fast hue-based colour image indexing algorithm. *Machine Graphics & Vision International Journal*, 11(2/3):285–295, 2002. 3

[10] S.-H. Cha and S. N. Srihari. On measuring the distance between histograms. *Pattern Recognition*, 35(6):1355–1370, 2002. 3, 4

[11] K.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Ordinal hyperplanes ranker with cost sensitivities for age estimation. In *Computer vision and pattern recognition (cvpr), 2011 ieee conference on*, pages 585–592. IEEE, 2011. 2

[12] J. Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968. 6

[13] M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in neural information processing systems*, pages 2292–2300, 2013. 2, 3, 6

[14] J. F. P. da Costa, H. Alonso, and J. S. Cardoso. The unimodal model for the classification of ordinal data. *Neural Networks*, 21(1):78–91, 2008. 2, 3

[15] J. Delon, J. Salomon, and A. Sobolevski. Fast transport optimization for monge costs on the circle. *SIAM Journal on Applied Mathematics*, 70(7):2239–2258, 2010. 3

[16] G. Divon and A. Tal. Viewpoint estimationinsights & model. *arXiv preprint arXiv:1807.01312*, 2018. 2

[17] C. Frogner, C. Zhang, H. Mobahi, M. Araya, and T. A. Poggio. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pages 2053–2061, 2015. 2, 3, 6

[18] V. Gulshan, L. Peng, M. Coram, M. C. Stumpe, D. Wu, A. Narayanaswamy, S. Venugopalan, K. Widner, T. Madams, J. 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. 1

[19] P. A. Gutierrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2016. 3

[20] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[21] P. J. Huber. Robust statistics. In *International Encyclopedia of Statistical Science*, pages 1248–1251. Springer, 2011. 2, 3

[22] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012. 8

[24] X. Liu, B. V. K. V. Kumar, C. Yang, Q. Tang, and J. You. Dependency-aware attention control for unconstrained face recognition with image sets. 2018. 4

[25] X. Liu, Y. Zou, Y. Song, C. Yang, J. You, and B. V. Kumar. Ordinal regression with neuron stick-breaking for medical diagnosis. In *European Conference on Computer Vision*, pages 335–344. Springer, 2018. 2, 3, 7, 8

[26] Y. Liu, A. W.-K. Kong, and C. K. Goh. Deep ordinal regression based on data relationship for small datasets. In *IJCAI*, pages 2372–2378, 2017. 3

[27] Z. Ma and S. Chen. A convex formulation for multiple ordinal output classification. *Pattern Recognition*, 86:73–84, 2019. 3

[28] S. Mahendran, H. Ali, and R. Vidal. A mixed classification-regression framework for 3d pose estimation from 2d images. *BMVC*, 2018. 2

[29] M. Martinez, M. Tapaswi, and R. Stiefelhagen. A closed-form gradient for the 1d earth movers distance for spectral deep learning on biological data. In *ICML 2016 Workshop on Computational Biology (CompBio@ ICML16)*, 2016. 3

[30] F. Massa, R. Marlet, and M. Aubry. Crafting a multi-task cnn for viewpoint estimation. *British Machine Vision Conference*, 2016. 2

[31] K. I. Mortensen, L. S. Churchman, J. A. Spudich, and H. Flyvbjerg. Optimized localization analysis for single-molecule tracking and super-resolution microscopy. *Nature methods*, 7(5):377, 2010. 3

[32] R. M. Nishikawa, C. E. Comstock, M. N. Linver, G. M. Newstead, V. Sandhir, and R. A. Schmidt. Agreement between radiologists interpretations of screening mammograms. In *International Workshop on Digital Mammography*, pages 3–10. Springer, 2016. 2

[33] Z. Niu, M. Zhou, L. Wang, X. Gao, and G. Hua. Ordinal regression with multiple output cnn for age estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4920–4928, 2016. 2, 6

[34] O. Pele and M. Werman. A linear time histogram metric for improved sift matching. In *European conference on computer vision*, pages 495–508. Springer, 2008. 3, 4

[35] G. Pereyra, G. Tucker, J. Chorowski, Ł. Kaiser, and G. Hinton. Regularizing neural networks by penalizing confident output distributions. *arXiv preprint arXiv:1701.06548*, 2017. 2, 3

[36] J. Rabin, J. Delon, and Y. Gousseau. A statistical approach to the matching of local features. *SIAM Journal on Imaging Sciences*, 2(3):931–958, 2009. 5

[37] V. Ratner, Y. Shoshan, and T. Kachman. Learning multiple non-mutually-exclusive tasks for improved classification of inherently ordered labels. *arXiv preprint arXiv:1805.11837*, 2018. 3, 6, 7, 8

[38] M. Raza, Z. Chen, S.-U. Rehman, P. Wang, and P. Bao. Appearance based pedestrians head pose and body orientation estimation using deep learning. *Neurocomputing*, 272:647–659, 2018. 1

[39] M. L. Rizzo and G. J. Székely. Energy distance. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8(1):27–38, 2016. 6

[40] Y. Rubner, C. Tomasi, and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International journal of computer vision*, 40(2):99–121, 2000. 2, 4

[41] L. Rüschendorf. The wasserstein distance and approximation theorems. *Probability Theory and Related Fields*, 70(1):117–129, 1985. 2

[42] A. J. Salazar, J. A. Romero, O. A. Bernal, A. P. Moreno, and S. C. Velasco. Reliability of the bi-rads final assessment categories and management recommendations in a telemammography context. *Journal of the American College of Radiology*, 14(5):686–692, 2017. 2

[43] B. Su and G. Hua. Order-preserving wasserstein distance for sequence matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2906–2914, 2017. 3

[44] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 2, 3, 4, 5

[45] R. Szeto and J. J. Corso. Click here: Human-localized keypoints as guidance. *arXiv preprint arXiv:1703.09859*, 2017. 3

[46] C. Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003. 3, 5, 6

[47] M. Werman, S. Peleg, R. Melter, and T. Y. Kong. Bipartite graph matching for points on a line or a circle. *Journal of Algorithms*, 7(2):277–284, 1986. 4, 5

[48] H. Zhao, Z. Wang, and P. Liu. The ordinal relation preserving binary codes. *Pattern Recognition*, 48(10):3169–3179, 2015. 3