Not All Samples are Trustworthy: Towards Deep Robust SVP Prediction

Qianqian Xu, *IEEE Senior Member,* Zhiyong Yang, Yangbangyan Jiang
Xiaochun Cao, *IEEE Senior Member,* Yuan Yao, and Qingming Huang*, *IEEE Fellow*

Abstract—In this paper, we study the problem of estimating subjective visual properties (SVP) for images, which is an emerging task in Computer Vision. Generally speaking, collecting SVP datasets involves a crowdsourcing process where annotations are obtained from a wide range of online users. Since the process is done without quality control, SVP datasets are known to suffer from noise. This leads to the issue that not all samples are trustworthy. Facing this problem, we need to develop robust models for learning SVP from noisy crowdsourced annotations. In this paper, we construct two general robust learning frameworks for this application. Specifically, in the first framework, we propose a probabilistic framework to explicitly model the sparse unreliable patterns that exist in the dataset. It is noteworthy that we then provide an alternative framework that could reformulate the sparse unreliable patterns as a "contraction" operation over the original loss function. The latter framework leverages not only efficient end-to-end training but also rigorous theoretical analyses. To apply these frameworks, we further provide two models as implementations of the frameworks, where the sparse noise parameters could be interpreted with the HodgeRank theory. Finally, extensive theoretical and empirical studies show the effectiveness of our proposed framework.

Index Terms—Subjective Visual Property (SVP); Robustness; Outlier Detection; Probabilistic Model.

1 Introduction

With the increasing popularity of human-centric computer vision (CV) applications, Subjective Visual Properties (SVP) prediction [11], [22], [29], as an emerging CV task of this kind, has attracted a substantial amount of attention in the community. Generally speaking, SVP measures a user's subjective perception and feeling, concerning a particular property in images/videos. As typical instances, estimating properties of consumer goods such as shininess of shoes [11] improves customer experiences on online shopping websites; and estimating interestingness [10] from images/videos would be helpful for media-sharing websites (e.g., YouTube).

In this task, the most critical issue lies in how to measure the strength of SVP perception coming from an individual. Traditional methods usually adopt absolute value to specify a rating from 1 to 5 (or, 1 to 10) to grade the property of a stimulus. For example, in image/video interestingness prediction, a score of 5 is adopted to represent the most interesting items, while a score of 1 is adopted to present the least ones. However, since these properties are subjective, different raters often exhibit different interpretations of the scales, which leads to the fact that the annotations of different people on the same sample can vary widely. Moreover, it is hard to concretely define the concept of scale (for example, what a score of 3 exactly means for an image), especially without any common reference point. Therefore, recent investigations turn their focus to an alternative approach that adopts pairwise comparisons. In these studies, an individual is simply asked to compare two stimuli simultaneously, and votes which one has the stronger property based on his/her perception. Therefore individual decision process in pairwise comparison becomes much easier to control than in the absolute value case, as the multiple-scale rating is reduced to a dichotomous choice. Consequently, adopting the pairwise comparison can no doubt yield more reliable feedback with

- *Q. Xu is with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, (e-mail: xuqianqian@ict.ac.cn).*
- *Z. Yang and Y. Jiang are with State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: yangzhiyong@iie.ac.cn, jiangyangbangyan@iie.ac.cn).*
- *X. Cao is with State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, also with Cyberspace Security Research Center, Peng Cheng Laboratory, Shenzhen 518055, China, and also with School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: caoxiaochun@iie.ac.cn).*
- *Y. Yao is with Department of Mathematics, and by courtesy, Computer Science and Engineering, Hong Kong University of Science and Technology, Hong Kong, (e-mail: yuany@ust.hk).*
- *Q. Huang is with the School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing 101408, China, also with the Key Laboratory of Big Data Mining and Knowledge Management (BDKM), University of Chinese Academy of Sciences, Beijing 101408, China, also with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China, and also with Peng Cheng Laboratory, Shenzhen 518055, China (e-mail: qmhuang@ucas.ac.cn).*
- **Corresponding author.*

less personal scale bias in practice. However, a shortcoming of pairwise comparison is that it often suffers from a larger sampling complexity than the absolute value annotations, since the number of pairs grows quadratically with the number of items to be ranked.

Facing such a huge demand of data samples, one solution is to collect annotations from a large, relatively open, and rapidly-evolving group of Internet users. Fortunately, this could be exactly realized by the so-called crowdsourcing [3] platforms (such as MTurk, InnoCentive, CrowdFlower, CrowdRank, and AllOurIdeas), where the annotation tasks are assigned to a wide spectrum of the raters hired in these platforms. In this way, resorting to crowdsourcing certainly relieves the burden of data collecting. *But are these collected samples always trustworthy?* The answer is, however, negative. Since the raters in this case often work in the absence of strong quality control, it is hard to guarantee the annotation quality in general [7]. The sources behind this issue could be concluded as the following cases. First of all, the raters are always impatient. Confronting a large number of annotations, the raters might probably finish his/her work in a hurry with random annotations. More seriously, the malicious users might even provide wrong answers deliberately to corrupt the system. Such contaminated decisions are useless and may hurt the total prediction performance heavily. This makes it inevitable to identify and remove such harmful annotations to achieve a robust SVP prediction result.

To address this issue, recent studies against outliers usually adopt the following two-stage framework: the first step is to perform a standard outlier detection process (e.g., majority voting) followed with the second step where a regression or learning to rank model is built. However, it has been pointed out that when pairwise local rankings are integrated into a global ranking, traditional outlier detection methods (e.g., majority voting) tend to ignore the outliers that can cause global inconsistency and yet are locally consistent [16]. To overcome this limitation, [11] proposes a more principled way to identify annotation outliers by formulating the SVP prediction task as a unified robust learning to rank problem, tackling both the outlier detection and SVP prediction tasks jointly. The common issue in this line of previous work is that they only restrict their focus on shallow models bearing limited representation power with low-level image features. Our goal in this paper is to leverage the strong representation power of deep neural networks to explore the SVP prediction issue from a deep perspective.

Under the croudsourcing scenario, employing deep learning models makes maintaining robustness becomes even more challenging. Deep learning is more or less vulnerable to contaminated data since the high complexity brings extra risks to overfit the noisy data [23], [12], [37], [39], [18], [30], [27]. Hence, we believe that how to guarantee the robustness is one of the biggest challenges for crowdsourced deep SVP prediction models. In this sense, we propose a deep robust model for learning SVP from crowdsourcing.

Specifically, on top of a Siamese-based architecture, we propose two general robust learning frameworks for this application. The first framework adopts explicit modeling of a set of sparse noises, which accounts for the unreliable patterns in the crowdsourced dataset. However, we find that, with the sparse parameters, it is hard to be trained in an end-to-end manner. This motivates us to derive another framework based on reformulating the previous framework. The new framework transforms the sparse parameters as a "contraction" operation over the original loss function, where not only could the sparse parameters be canceled completely, but the robustness of the framework can also be justified by rigorous theoretical analyses.

As a summary, we list our main contributions as follows:

(C1) Two general frameworks are proposed in this paper for robust crowdsourced SVP ranking. To the best of our knowledge, our frameworks offer the first attempt to carry out the prediction procedure with the automatic detection of sparse outliers from a deep perspective.

(C2) In the first framework, we propose a general probabilistic framework to explicitly model unreliable patterns as sparse learnable noises. In the second framework, we provide a much simpler reformulation that simultaneously leverages end-to-end training and theoretical justifications.

(C3) We provide two implementations of the frameworks, where the sparse learnable noises could explicitly interpret the inconsistent rankings in the dataset.

This paper is an extension of our conference paper [35], where we proposed a unified robust deep learning framework for pairwise SVP prediction, along with two instantiations. The novel results in this long version include the following. First Thm.1 provides a unified framework to reformulate the original models. Thm.2 reveals how robustness explicitly takes place in the proposed framework. Thm.3 shows that the robustness of the proposed framework further leverages improved generalization ability compared with the vanilla model. Finally, based on Thm.4 and Col.1-2, we propose two concrete models as implementations of the proposed frameworks with an interpretable formulation of the sparse learnable noises. Moreover, the new version also includes a substantial amount of experimental extensions, which includes systematic ablation studies and sensitivity studies shown in Sec.6.

The rest of this paper is organized as follows. Sec.2 contains a review of related work. Then we systemat-

ically introduce our problem setup in Sec.3. In Sec.4, we elaborate the proposed frameworks. Two concrete implementations are then proposed in Sec.5. Next, extensive experimental validations based on three real-world crowdsourced datasets are demonstrated in Sec.6. Finally, Sec.7 presents the conclusive remarks for this paper.

2 Related Work

2.1 Subjective visual properties

Subjective visual property prediction has gained rising attention in the last several years. The related work involves a large variety of computer vision problems, ranging from image/video interestingness [10], memorability [19], [29], to quality of experience prediction [33], [1]. In this paper we restrict our focus to the pairwise SVP prediction task, where the subjective visual properties are often called relative attributes [36], [22]. The historical studies treat this task as a learning-to-rank problem. The main idea is to use ordered pairs of training images to train a ranking function. Specifically, a set of pairs ordered according to their perceived property strength is obtained from human annotators, and a ranking function that preserves those orderings is learned. Given a new image pair, the ranker indicates which image has the property more. Popular approaches to learning-to-rank for pairwise data include RankSVM [20], RankBoost [8], and RankNet [4], etc. However, these methods are not a natural fit facing the crowdsourced outliers. Seeing this issue, [11] proposes a unified robust learning to rank (URLR) framework to conduct outlier detection and learning to rank in a joint manner. Though aimed at the same purpose, our study differs significantly from this literature in the following sense. Firstly, we propose a general framework based on a probabilistic model, while [11] only considers a special case of our framework with the squared loss. Secondly, our study also unifies outliers detection and deep learning models, while [11] is only restricted to shallow features. Last but not least, we provide theoretical analysis on how and why robustness takes place in the methodology.

2.2 Learning with noisy data

Learning from noisy data has gained increasing attention in recent years. Traditionally, such methods include Majority voting, M-estimator [15], Transitivity Satisfaction Rate (TSR) [5], Huber-LASSO [33], and Least Trimmed Squares (LTS) [34], etc. Recently, there is a wave to explore robust methods in the context of deep learning. Generally speaking, these studies fall into four brunches: (**I**) probabilistic graphical models where the noisy patterns are modeled as latent variables; (**II**) progressive and self-paced models, where easy and clean examples are learned first while the hard/noisy labels are progressively learned; (**III**) loss-correction methods, where the loss function is corrected iteratively; (**IV**) network architecture based method, where the noisy patterns are modeled with specifically designed modules. In the following, we enumerate some typical examples for each type of method. For (**I**): in [37], a novel quality variable is introduced to the deep learning models, so as to measure the trustworthiness of noisy labels automatically; [30] proposes a novel framework for training deep convolution neural networks from noisy labeled datasets, where an undirected graphical model is constructed to model the relationship between noisy and clean labels. For (**II**), [12] proposes a Progressive Stochastic Learning framework to learn adaptively from clean to noisy labels; [17] proposes a self-paced unsupervised deep learning method, where noisy instances are not activated in the training process until the model becomes sufficiently sophisticated. For (**III**), [27] forms a loss correction model: it proposes two procedures for loss correction that are agnostic to both application domain and network architecture. For (**IV**), [18] proposes a simple but effective model to formulate for label noise within deep neural networks, where a softmax layer is employed to explicitly model the label noise statistics. Meanwhile, there are also some efforts on constructing deep robust models for specific tasks and applications: [23] proposes a method to learn from weak and noisy labels for Semantic Segmentation; [39] proposes a deep robust unsupervised method for saliency detection, etc.

Compared with these recent achievements, our work differs significantly in the sense that: a) We provide the first trial to explore the deep robust learning problem in the context of crowdsourced SVP learning. b) We adopt a pairwise learning framework, whereas the existing work all adopts instance-wise frameworks. c) Our proposed frameworks enjoy rigorous theoretical guarantees concerning both robustness and improved generalization ability.

3 Problem Setup

Our goal in this paper is two-fold:

- **(a)** We aim to learn a deep SVP prediction model from a set of sparse and noisy pairwise comparison labels. Specifically the ranking patterns should be preserved.
- **(b)** To guarantee the quality of the model, we expect that all the noisy annotations could be detected and removed along with the training process.

With our goal claimed, we now step further to a detailed formulation of the data structure adopted in this paper. In the SVP prediction task, we are given a pool with n training images and a set of SVPs. In addition, for each SVP, we are given a set of pairwise comparison labels. Each sample in this training data contains a pairwise comparison result for two images

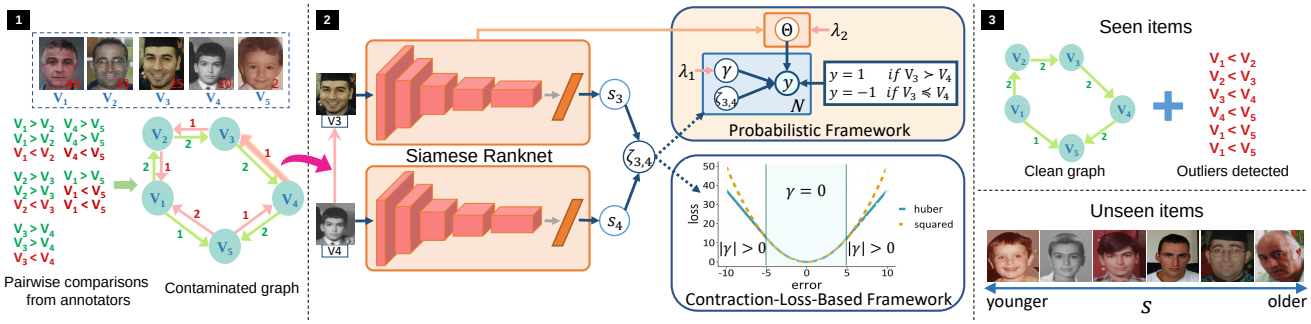


Fig. 1: Overview of our approach. (1) Constructing a comparison graph from the crowdsourcing annotations, which is contaminated with outlier labels. (2) We propose a generalized deep probabilistic framework, where an outlier parameter γ is learned along with the network parameters Θ. Moreover, we propose a reformulated framework, which could serve the same purpose even without γ. (3) Our framework will output a clean graph on the training set, where contaminated annotations are eliminated. Furthermore, our model could predict a rank-preserved score for each unseen instance. Best viewed in color.

TABLE 1: Comparisons of the proposed frameworks.

Formulation	Outlier Detection	Extra Parameters	End-to-end Training Strategy
$\min_{\Theta,\boldsymbol{\gamma}} \frac{1}{m} \sum_{(i,j)\in\mathcal{D}} \tilde{f}(\omega_{i,j}+\gamma_{i_1,i_2}^{y_{i,j}}) + \lambda_1\|\boldsymbol{\gamma}\|_1 + \lambda_2 \sum_{\theta\in\Theta}\theta^2$	✔	$\boldsymbol{\gamma}$	×
$\min_{\Theta} \sum_{(i,j)\in\mathcal{D}} \tilde{\ell}(\omega_{i,j},\lambda_1) + \lambda_2 \cdot \sum_{\theta\in\Theta}\theta^2$	✔	×	✔

TABLE 2: Notations and descriptions.

Notation	Description
(i,j)	a shorthand for $(i_1,i_2,y_{i,j})$
$\boldsymbol{x}_i$	the raw feature for image i
$\boldsymbol{X}$	the set of all features in the training data
Θ	the set of all the parameters in a network
$s(\boldsymbol{x}_i)$	the predicted score of image i
$\gamma_{i_1,i_2}^{y_{i,j}}$	the noise parameter for the comparison (i,j)
$\boldsymbol{\gamma}$	the set of all noises$\{\gamma_{i_1,i_2}^{y_{i,j}}\}_{(i,j)}$
$n_{i_1,i_2}^{y_{i,j}}$	the number of times (i_1,i_2) is labeled as $y_{i,j}$
ζ_{i_1,i_2}	$s(\boldsymbol{x}_{i_1}) - s(\boldsymbol{x}_{i_2})$
$\omega\left(y_{i,j},\zeta_{i_1,i_2}\right)$	the general form of a predictive function
$\omega_{i,j}$	the shorthand of $\omega\left(y_{i,j},\zeta_{i_1,i_2}\right)$
$\omega_{i,j}+\gamma_{i_1,i_2}^{y_{i,j}}$	the noisy form of the predicted score
f	the conditional p.d.f. of the labels
Range	Range(f) is the range of f
$\mathcal{LIP}$	$\mathcal{LIP}(f)$ is the Lipschitz constant of f
$\tilde{f}$	the original loss function
$\tilde{\ell}(\cdot,\lambda_1)$	the reformulated loss function
$\tilde{f}'$	the derivative of $\tilde{f}$
ψ	$\tilde{f}'^{-1}$
ϕ	$\tilde{f}\circ\psi$

from a given user. Given such a sample for image pair i labeled by user j, we denote (i,j) as the corresponding id. Furthermore, we denote the ids of the two images as i_1 and i_2, the corresponding image pair as $(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2})$ and the annotation as $y_{i,j}$. From a global view, such pairwise samples can be represented by a directed multi-graph where multiple edges could be found between two vertexes. Mathematically, we denote the graph as $\mathcal{G}=(\mathcal{V},\mathcal{E})$. $\mathcal{V}$ is the set of vertexes that contains all the distinct image items that occurred in the comparisons. $\mathcal{E}$ is the set of comparison edges. For a specific user with id j and a specific comparison pair i defined on two vertexes i_1 and i_2, if the user believes that i_1 holds a stronger/weaker presence of the SVP, we then have an edge $(i_1,i_2,j)/(i_2,i_1,j)$, respectively. Equivalently we also denote this relation as $i_1 \overset{j}{\succ} i_2 / i_2 \overset{j}{\succ} i_1$. Since multiple users take part in the annotation process, it is natural to observe multi-edges between two vertexes. Now we could denote the labeling results as a function $\mathcal{Y}:\mathcal{E}\to\{-1,1\}$. For a given pair i and a rater j who annotates this pair, the corresponding label is denoted as $y_{i,j}$, which is defined as:

$$\begin{cases} y_{i,j} = \quad 1, & (i_1,i_2,j)\in\mathcal{E}; \\ y_{i,j} = -1, & (i_2,i_1,j)\in\mathcal{E}. \end{cases} \tag{1}$$

Now we present an example of the defined comparison graph. See step 1 in Fig.1. In this figure, the SVP in question is the age of the humans in the images. Suppose we have 5 images with ground-truth ages (marked with red in the lower right corner of each image), we then have $\mathcal{V}=\{1,2,\cdots,5\}$. Furthermore, we have three users with id 1, 2, 3 who take part in the annotation. According to the labeling results shown in the lower left side, we have $\mathcal{E}=\{(1,2,1),(1,2,2),(2,1,3),\cdots,(1,5,1),(5,1,2),(5,1,3)\}$. As shown in this example, we would be most likely to observe both $i_1\succ i_2$ and $i_2\succ i_1$ for a specific pair i. This is mainly caused by the malicious users who provide erroneous labels. For example, for vertexes 1 and 2, the edge $(2,1,3)$ is obviously an abnormal

annotation. This phenomenon coincides with the arguments in the introduction section.

With the above definitions and explanations, we are ready to introduce the input and output of our proposed model.

Input. The input of our deep model is the defined multi-graph $\mathcal{G}$ along with the image items, where each time a specific edge is fed to the network.

Output. As will be seen in the next paragraph, our model will output the relative score $s(\boldsymbol{x}_{i_1})$ and $s(\boldsymbol{x}_{i_2})$ of the image pair along with an outlier parameter which could automatically remove the abnormal directions on $\mathcal{G}$. Note that learning $s(\boldsymbol{x}_{i_1})$ and $s(\boldsymbol{x}_{i_2})$ directly achieves our goal **(a)**, while detecting and removing outlier directions on the graph directly achieves goal **(b)**.

Architecture. Given the problem definition, next we elaborate the architecture we adopted in this work, see Fig.1 as an overview. According to step 2 in Fig.1, we employ a deep Siamese [6], [26] convolutional neural network as the ranking model to learn the relative scores for image pairs. In this model, the input is an edge in the graph $\mathcal{G}$ together with the image pair $(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2})$. Each branch of the network is fed with an image and outputs the corresponding scores $s(\boldsymbol{x}_{i_1})$ and $s(\boldsymbol{x}_{i_2})$. Then we propose a robust probabilistic model based on the difference between the scores. As a note for the network architecture, we choose an existing popular CNN architecture, ResNet-50 [13], as the backbone of the Siamese network. Such residual network is equipped with shortcut connections, bringing in promising performance in image tasks.

4 A Tale of Two Frameworks

In this section, we propose two frameworks for robust crowd-sourced ranking. The first framework follows a straight-forward intuition. Specifically, we explicitly employ a set of learnable sparse additive noises $\boldsymbol{\gamma}$ in a general probabilistic model. With such noises included, we expect to detect the outliers in the dataset from the samples where the corresponding $\boldsymbol{\gamma}$ is non-zero. However, with a close look at the details, we found that it is hard to perform end-to-end training with this framework. Moreover, it is unclear how robustness takes place in a general sense. **Surprisingly, we find that the sparse noises in the first framework could be canceled with a simple "contraction" operation over the loss function.** This result leads us to our second framework, where we can not only perform end-to-end training but also provide a theoretical insight into how generalization is improved with this robust learning framework. Moreover, in the new framework, we could find out the outliers as well, even without the effort of $\boldsymbol{\gamma}$. The readers are referred to Tab.2 for a quick overview of all the frequently appeared notations. See Tab.1 as a summary of the comparisons.

In Sec.4.1, we will present the first framework, while in Sec.4.2 we will present the second framework.

4.1 Robust Models with Learnable Sparse Additive Noises: A Straight-forward Formulation

Recall the network architecture given in the previous section, we are ready to elaborate on a novel probabilistic framework to simultaneously prune the outliers and learn the network parameters for SVP prediction. In our model, noisy annotations are treated as a mixture of reliable patterns and outlier patterns. More precisely, to guarantee the performance of the whole model, we expect $s(\boldsymbol{x}_{i_1}), s(\boldsymbol{x}_{i_2})$, i.e., the scores returned by the network to capture the reliable patterns in the labels. Meanwhile, we introduce an outlier probe variable $\gamma_{i_1,i_2}^{y_{i,j}}$ to model the noisy nature of the annotations. During the training process, our prediction is an additive mixture of the reliable score and the outlier probe. Given the clarification above, our next step is to propose a probabilistic model of the generation process of the labels based on the outlier noise parameter $\boldsymbol{\gamma}$, the network parameters Θ, and the predicted scores $s(\cdot)$. More precisely, we model the conditional distribution of the annotations along with the prior distribution of $\boldsymbol{\gamma}$ and Θ in the following form:

$$y_{i,j} \mid \boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, \Theta, \gamma_{i_1,i_2}^{y_{i,j}} \overset{i.i.d}{\sim} f(\omega\left(y_{i,j}, \zeta_{i_1,i_2}\right) + \gamma_{i_1,i_2}^{y_{i,j}}),$$

$$\gamma_{i_1,i_2}^{y_{i,j}} \mid \lambda_1 \overset{i.i.d}{\sim} h(\gamma_{i_1,i_2}^{y_{i,j}}, \lambda_1), \quad \Theta \mid \lambda_2 \sim g(\Theta, \lambda_2).$$

1) $\zeta_{i_1,i_2} = s(\boldsymbol{x}_{i_1}) - s(\boldsymbol{x}_{i_2})$ is the relative score of the annotation, which will be directly learned from the deep learning model with the parameter set Θ. As mentioned above, ζ_{i_1,i_2} is expected to model the reliable pattern in the annotations. The prior distribution of Θ is assumed to be associated with a p.d.f. (probability density function) $p(\Theta \mid \lambda_2) = g(\Theta, \lambda_2)$ (λ_2 is a predefined hyperparameter), which is denoted as g in short.
2) $\gamma_{i_1,i_2}^{y_{i,j}}$ is the outlier probe which induces unreliability. Since only outliers have a nonzero $\gamma_{i_1,i_2}^{y_{i,j}}$, we model the randomness of $\gamma_{i_1,i_2}^{y_{i,j}}$ with an i.i.d sparsity-inducing prior distribution (e.g., Laplacian distribution) with the p.d.f. being $p(\gamma_{i_1,i_2}^{y_{i,j}}|\lambda_1) = h(\gamma_{i_1,i_2}^{y_{i,j}}, \lambda_1)$ (λ_1 denotes the hyperparameter), which is denoted as h_{ij} in short.
3) As we have mentioned above, the noisy prediction $\omega\left(y_{i,j}, \zeta_{i_1,i_2}\right) + \gamma_{i_1,i_2}^{y_{i,j}}$ is an additive mixture of the reliable score and outlier parameter.
4) $f(\omega\left(y_{i,j}, \zeta_{i_1,i_2}\right) + \gamma_{i_1,i_2}^{y_{i,j}})$ is the conditional p.d.f. of the labels, which is denoted as f_{ij} in short.

Let $\boldsymbol{\gamma} = \{\gamma_{i_1,i_2}^{y_{i,j}}\}_{(i_1,i_2,j)\in\mathcal{E}}$, $\boldsymbol{y} = \{y_{i,j}\}_{(i_1,i_2,j)\in\mathcal{E}}$. Now our next step is to construct a loss function for this probabilistic framework. According to the Maximum A Posterior (MAP) rule in statistics, a reasonable solution of the parameters should have a large posterior

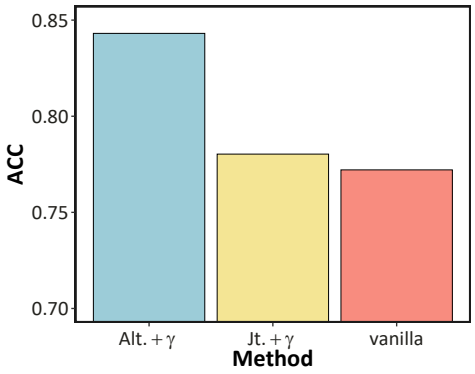


Fig. 2: **Comparison of the training strategies over the Human Age Dataset.** Here "Alt.+γ" represents the test set performance from the alternative training of the robust learning framework, "Jt.+γ" represents the result from the joint training of the robust learning framework, "vanilla" represents the result from vanilla deep learning model without robust learning. We use **Model A** proposed in Sec.5 as an example.

probability $P(\Theta, \boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{X}, \lambda_1, \lambda_2)$. In other words, with high probability, the parameters ($\boldsymbol{\gamma}$ *and* Θ *in our model*) should be observed after seeing the data ($\boldsymbol{y}$ *and* $\boldsymbol{X}$ *in our model*) and the predefined hyperparameters (λ_1, λ_2). This motivates us to *maximize* the posterior probability in our objective function. Furthermore, to simplify the calculation of the derivatives, we adopt an equivalent form where the negative log posterior probability is *minimized*:

$$\min_{\Theta,\boldsymbol{\gamma}} -\log\left(P(\Theta, \boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{X}, \lambda_1, \lambda_2)\right).$$

The Bayesian rule implies that our loss function could be simplified as:

$$\min_{\Theta,\boldsymbol{\gamma}} \sum_{(i,j)\in\mathcal{D}} -\left(\log(f_{ij}) + \log(h_{ij})\right) - \log(g). \tag{2}$$

More specifically, to leverage the sparsity of $\gamma_{i_1,i_2}^{y_{i,j}}$, we narrow our focus to a restricted family of models where we set $h_{ij} = \exp(-\lambda_1 \cdot |\gamma_{i_1,i_2}^{y_{i,j}}|)$, $g = \prod_{\theta\in\Theta} \exp(\lambda_2 \cdot \theta^2)$ and $\log(f_{i,j}) = \tilde{f}\left(\omega(y_{i,j}, \zeta_{i_1,i_2}) + \gamma_{i_1,i_2}^{y_{i,j}}\right)$. The corresponding optimization problem becomes:

$$(Org)\ \min_{\Theta,\boldsymbol{\gamma}} \frac{1}{m} \sum_{(i,j)\in\mathcal{D}} \tilde{f}(\omega_{i,j} + \gamma_{i_1,i_2}^{y_{i,j}}) + \lambda_1 \|\boldsymbol{\gamma}\|_1 + \lambda_2 \sum_{\theta\in\Theta} \theta^2,$$

where $\omega_{i,j}$ stands for $\omega\left(y_{i,j}, \zeta_{i_1,i_2}\right)$, $\|\boldsymbol{\gamma}\|_1 = \sum_{(i,j)\in\mathcal{D}} n_{i_1,i_2}^{y_{i,j}} |\gamma_{i_1,i_2}^{y_{i,j}}|$ and $n_{i_1,i_2}^{y_{i,j}} = |(i_1, i_2, y_{i,j})|$ (the number of times (i_1, i_2) is labeled as $y_{i,j}$).

Optimization. Note that it is hard to train this model in an end-to-end manner. The major source for this hardness is that $\omega_{i,j}$ and $\gamma_{i_1,i_2}^{y_{i,j}}$ are combined together with an addition operation $\omega_{i,j} + \gamma_{i_1,i_2}^{y_{i,j}}$. This means that a local optimum could be achieved by either optimizing $\gamma_{i_1,i_2}^{y_{i,j}}$ or $\omega_{i,j}$ until $\omega_{i,j} + \gamma_{i_1,i_2}^{y_{i,j}}$ becomes a reasonable prediction. In other words, $\gamma_{i_1,i_2}^{y_{i,j}}$ and $\omega_{i,j}$ have to compete with each other to gain their own gradient information. This makes the algorithm unable to tell apart what is the pattern and what is the noise. In this sense, we have to adopt an alternating optimization strategy to train this model. The network is firstly trained with several epochs, $\boldsymbol{\gamma}$ is then trained after that, when we have got good estimations of $\omega_{i,j}$. *The benefit from alternating training could be seen from Fig.2.* Here, we take the Human Age dataset as an example. The alternating training scheme provides a significant improvement compared with the vanilla network. However, if joint training is adopted, one can only observe a tiny improvement. The readers are referred to the appendix for a detailed algorithm for this training strategy applied to the two concrete models proposed in Sec.5.

Outlier Detection. As another matter of fact, detecting outliers in this framework relies on the estimation of the noises in $\boldsymbol{\gamma}$. Typically, we could filter out the outliers with top magnitude of $\boldsymbol{\gamma}$ and keep the samples whose corresponding $\gamma_{i_1,i_2}^{y_{i,j}}$ is zero or small in magnitude. Though this is a straight-forward strategy to perform outlier detection, it requires an extra set of $O(m)$ parameters compared with the vanilla deep model.

Above all, we have seen that the limitations of this straight-forward framework come from the sparse variable $\boldsymbol{\gamma}$. It then poses the question that *can we entirely cancel* $\boldsymbol{\gamma}$ *out from framework*, such that it could behave like the vanilla deep model. Fortunately, in the next subsection, we will provide a positive answer to this question.

4.2 Robust Models via Contraction: A Simple Reformulation

In this subsection, we will provide a much simpler reformulation of the original framework proposed in the last subsection. As a surprising matter of fact, the new framework could address all the issues mentioned above. **More importantly, the reformulation also makes it possible to prove that the proposed robust learning framework could also improve the generalization ability of the vanilla deep learning model**.

First of all, the following theorem provides the reformulation of the original problem.

Theorem 1 (Unified Reformulation of the Sparse Additive Noise Model). *Considering the model formulation in Eq.(2), we set* $h_{ij} = \exp(-\lambda_1 \cdot |\gamma_{i_1,i_2}^{y_{i,j}}|)$, $g = \prod_{\theta\in\Theta} \exp(\lambda_2 \cdot \theta^2)$ *and* $\log(f_{i,j}) = \tilde{f}\left(\omega(y_{i,j}, \zeta_{i_1,i_2}) + \gamma_{i_1,i_2}^{y_{i,j}}\right)$, *where* $\tilde{f}$ *is a strictly convex function such that* $\tilde{f}' = \frac{d\tilde{f}(x)}{dx}$ *is invertible, continuous and strictly increasing, and* $\mathsf{Range}(\omega) = \mathbb{R}$. *For all* λ_1 *such that at least one element in* $\{-\lambda_1, \lambda_1\}$ *belongs to* $\mathsf{Range}(\tilde{f})$, *the original problem shares the same solution set with the*

following problem (Re)*:*

$$(Re)\quad \underset{\Theta}{\operatorname{argmin}} \sum_{(i,j)\in\mathcal{D}} \tilde{\ell}(\omega_{i,j},\lambda_1)+\lambda_2\cdot\sum_{\theta\in\Theta}\theta^2.$$

If we further define $\psi(x)=\tilde{f}'^{-1}(x), \phi(\cdot)=\tilde{f}\left(\psi(\cdot)\right)$ *and* $\omega_{i,j}=\omega\left(y_{i,j},\zeta_{i_1,i_2}\right)$*, then the following facts hold:*

(a) $\tilde{\ell}(\omega_{i,j},\lambda_1)$ *could be expressed as:*

(1) If $[-\lambda,\lambda]\subseteq \mathsf{Range}(\tilde{f}')$*,*

$$\begin{cases}\phi(-\lambda_1)+\lambda_1\cdot|\omega_{i,j}-\psi(-\lambda_1)|, & \omega_{i,j}<\psi(-\lambda_1),\\ \phi(\lambda_1)+\lambda_1\cdot|\omega_{i,j}-\psi(\lambda_1)|, & \omega_{i,j}>\psi(\lambda_1),\\ \tilde{f}(\omega_{i,j}), & \textit{otherwise},\end{cases}$$

(2) If $-\lambda\notin \mathsf{Range}(\tilde{f}')$*,*

$$\begin{cases}\tilde{f}(\omega_{i,j}), & \omega_{i,j}\le\psi(\lambda_1),\\ \phi(\lambda_1)+\lambda_1\cdot|\omega_{i,j}-\psi(\lambda_1)|, & \omega_{i,j}>\psi(\lambda_1).\end{cases}$$

(3) If $\lambda\notin \mathsf{Range}(\tilde{f}')$*,*

$$\begin{cases}\phi(-\lambda_1)+\lambda_1\cdot|\omega_{i,j}-\psi(-\lambda_1)|, & \omega_{i,j}<\psi(-\lambda_1),\\ \tilde{f}(\omega_{i,j}), & \omega_{i,j}\ge\psi(-\lambda_1),\end{cases}$$

(b) Given any feasible Θ*, the partial optimal solution* $\gamma_{i_1,i_2}^{y_{i,j}\star}$ *is:*

(1) If $[-\lambda,\lambda]\subseteq \mathsf{Range}(\tilde{f}')$*,*

$$\begin{cases}\psi(-\lambda_1)-\omega_{i,j}, & \omega_{i,j}<\psi(-\lambda_1),\\ 0, & \omega_{i,j}\in[\psi(-\lambda_1),\psi(\lambda_1)],\\ \omega_{i,j}-\psi(\lambda_1), & \omega_{i,j}>\psi(\lambda_1).\end{cases}$$

(2) If $-\lambda\notin \mathsf{Range}(\tilde{f}')$*,*

$$\begin{cases}0, & \omega_{i,j}\le\psi(\lambda_1),\\ \omega_{i,j}-\psi(\lambda_1), & \omega_{i,j}>\psi(\lambda_1).\end{cases}$$

(3) If $\lambda\notin \mathsf{Range}(\tilde{f}')$*,*

$$\begin{cases}\psi(-\lambda_1)-\omega_{i,j}, & \omega_{i,j}<\psi(-\lambda_1),\\ 0, & \omega_{i,j}\ge\psi(-\lambda_1).\end{cases}$$

Part (a) of Thm.1 shows that we can completely cancel out $\boldsymbol{\gamma}$ from the original problem with a reformulated loss, given mild conditions. The resulting loss provides a "contraction" transformation of the original loss. More specifically, it remains the same as the original loss when the magnitude of the gradient is moderate, while it behaves like a linear function when the magnitude of the gradient is larger than a predefined value. Part (b) of Thm.1 shows that the sparse parameter $\boldsymbol{\gamma}$ should also be recovered from the simple reformulated loss. In fact $\gamma_{i_1,i_2}^{y_{i,j}}$ is nonzero only if the reformulated loss function becomes linear. Moreover, the magnitude of nonzero $\gamma_{i_1,i_2}^{y_{i,j}}$s roughly account for the residual $|\omega_{i,j}-\psi(x)|$, and thus accounts for $|\tilde{f}'(\omega_{i,j})-\tilde{\ell}'(\omega_{i,j})|$, which is the difference between the gradients before and after the "contraction" transformation.

Since $\gamma_{i_1,i_2}^{y_{i,j}}$ no longer exists in the new framework, we can now train the models in an end-to-end manner easily.

With the basic properties clarified, we now take a step further to explore what makes this framework really robust against heavy noises. Our discussion is started with the following theorem.

Theorem 2 (Robustness of the Loss function). *Under the assumptions of Thm.1,* $\tilde{\ell}$ *has the following properties:*

(1) For all

$$\omega_{i,j}\in\mathcal{H}=\left\{\omega_{i,j}:\omega_{i,j}<\psi(-\lambda_1)\ \textit{or}\ \omega_{i,j}>\psi(\lambda_1)\right\},$$

we have:

$$\tilde{\ell}(\omega_{i,j},\lambda_1)<\tilde{f}(\omega_{i,j})$$

(2) $|\tilde{\ell}(\omega_{i,j},\lambda_1)-\tilde{\ell}(\omega'_{i,j},\lambda_1)|\le|\tilde{f}(\omega_{i,j})-\tilde{f}(\omega'_{i,j})|$*, with inequality holds strictly if at least element in* $\{\omega_{i,j},\omega'_{i,j}\}$ *belongs to* $\mathcal{H}$*.*

(3) $\mathcal{LIP}(\tilde{\ell}(\cdot,\lambda_1))=\lambda_1$*,* $\mathcal{LIP}(\tilde{f})>\lambda_1$*, where* $\mathcal{LIP}(\tilde{\ell}(\cdot,\lambda_1))$ *stands for the Lipschitz constant of* $\tilde{\ell}$ *w.r.t.* $\omega_{i,j}$*.*

(1) and (2) in the above theorem show that, compared with $\tilde{f}$, $\tilde{\ell}$ is less sensitive toward perturbations coming from $\omega_{i,j}$, i.e., the prediction function of the label. This reveals the major source of robustness.

Moreover, Part (3) of this theorem shows that $\tilde{\ell}$ could as well shrink the original Lipschitz constant of $\tilde{f}$. Besides improving the robustness, this also leads to improvement of the generalization ability. Next, we will investigate how such improvement takes place.

In this paper, given a hypothesis space $\mathcal{F}$, we will use the Rademacher complexity to derive the generalization bounds. A formal definition of the Rademacher complexity is as follows.

Definition 1 (Rademacher Complexity). *The empirical Rademacher complexity over a dataset* $\mathcal{S}=\{\boldsymbol{z}_i\}_{i=1}^m$*, and a hypothesis space* $\mathcal{F}$ *is defined as:*

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})=\frac{1}{m}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^m\sigma_i f(\boldsymbol{z}_i)\right],$$

where for $i=1,2,\cdots,N_C$*,* $\sigma_1,\cdots,\sigma_m$ *are i.i.d Rademacher random variables. The population version of the Rademacher Complexity is defined as* $\mathfrak{R}_{\mathcal{S}}(\mathcal{F})=\mathbb{E}_{\mathcal{S}}\left[\hat{\mathfrak{R}}_{\mathcal{S}}(\mathcal{F})\right]$*.*

For the ranking problem in our paper, we are interested in the following three hypothesis spaces. First denote $\mathcal{F}$ as the hypothesis space where the scoring function $s(\cdot)$ is chosen. Then denote

$$\tilde{\mathcal{F}}=\left\{\omega:\boldsymbol{z}=(\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2},y_{i,j})\to\omega\left(y_{i,j},\zeta_{i_1,i_2}\right),s(\cdot)\in\mathcal{F}\right\},$$

as the hypothesis space for the prediction function $\omega_{i,j}$. Finally,

$$\tilde{\ell}_{\lambda_1}\circ\tilde{\mathcal{F}}=\left\{\tilde{\ell}(\omega,\lambda_1),\ \omega\in\tilde{\mathcal{F}}\right\}$$

is defined as the hypothesis space with the loss function $\tilde{\ell}$ composed on top of $\omega_{i,j}$. Obviously, from the Talagrand Contraction Lemma [24, Lem.5.7], if ω is 1-Lipschitz continuous w.r.t. ζ_{i_1,i_2}, we have:

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\ell}_{\lambda_1} \circ \tilde{\mathcal{F}}) \leq \lambda_1 \cdot \hat{\mathfrak{R}}_{\mathcal{S}}(\tilde{\mathcal{F}}). \tag{3}$$

However, if we replace $\tilde{\ell}$ with $\tilde{f}$, the result turns out to be:

$$\hat{\mathfrak{R}}_{\mathcal{S}}(\tilde{f}' \circ \tilde{\mathcal{F}}) \leq \mathcal{LIP}(\tilde{f}) \cdot \hat{\mathfrak{R}}_{\mathcal{S}}(\tilde{f}' \circ \tilde{\mathcal{F}}).$$

Since $\mathcal{LIP}(\tilde{f}) > \lambda_1$ whenever $\lambda_1 \in \text{Range}(\tilde{f}')$, this induces improved generalization based on the following theorem.

Theorem 3 (Improved Generalization from Robustness). *Let $\mathcal{F}$ be the hypothesis space for the real scoring function $s(\cdot)$, such that $\zeta_{i_1,i_2} = s(\boldsymbol{x}_{i_1}) - s(\boldsymbol{x}_{i_2})$. Define the i.i.d training sample as $\mathcal{S} = \{(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, y_{i,j})\}_{i_1,i_2,j}$, and define two partial sets as $\mathcal{S}_1 = \{(\boldsymbol{x}_{i_1}, y_{i,j})\}_{i_1,j}$, and $\mathcal{S}_2 = \{(\boldsymbol{x}_{i_2}, y_{i,j})\}_{i_2,j}$. Moreover, let*

$$\mathfrak{R}_{\mathcal{S}_1}(\mathcal{F}) = \mathbb{E}_{\mathcal{S}}(\hat{\mathfrak{R}}_{\mathcal{S}_1}(\mathcal{F})),\ \mathfrak{R}_{\mathcal{S}_2}(\mathcal{F}) = \mathbb{E}_{\mathcal{S}}(\hat{\mathfrak{R}}_{\mathcal{S}_2}(\mathcal{F})),$$
$$\mathfrak{R}_{1+2} = \mathfrak{R}_{\mathcal{S}_1}(\mathcal{F}) + \mathfrak{R}_{\mathcal{S}_2}(\mathcal{F}),\ \hat{\mathfrak{R}}_{1+2} = \hat{\mathfrak{R}}_{\mathcal{S}_1}(\mathcal{F}) + \hat{\mathfrak{R}}_{\mathcal{S}_2}(\mathcal{F})$$
$$\hat{\mathcal{R}}_{\mathcal{S},\lambda} = \frac{1}{m} \sum_{(i,j)\in\mathcal{D}} \tilde{\ell}(\omega_{i,j}, \lambda),\ \mathcal{R}(s) = \mathbb{E}_{\mathcal{S}}[\sum_{(i,j)\in\mathcal{D}} \mathbf{1}_{y_{i,j}\cdot(\zeta_{i_1,i_2})\leq 0}]$$

The following two results about the generalization ability for $\tilde{\ell}$ based ERM hold:

1) *Under the assumption of Thm.1 if $\omega_{i,j} = y_{i,j} \cdot \zeta_{i_1,i_2}$ and $\mathbf{1}_{x\leq 0} \leq \tilde{\ell}(x, \lambda_1)$, $\text{Range}(\tilde{\ell}) = [0, B]$, with a fixed $\lambda_1 > 0$, for any $\delta \in (0,1)$, the following facts hold for all $s \in \mathcal{F}$ with probability at least $1-\delta$ over the choice of the training sample $\mathcal{S}$:*

$$\mathcal{R}(s) \leq \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) + 2 \cdot \lambda_1 \cdot \mathfrak{R}_{1+2} + B \cdot \sqrt{\frac{\log(1/\delta)}{2m}}$$

$$\mathcal{R}(s) \leq \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) + 2 \cdot \lambda_1 \cdot \hat{\mathfrak{R}}_{1+2} + 3B \cdot \sqrt{\frac{\log(2/\delta)}{2m}}$$

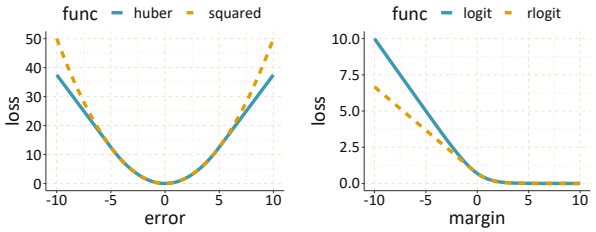
2) *Under the same assumption in (1), $\forall \delta \in (0,1)$ the following inequalities hold for all $\lambda_1 \in [r_1, r_2)$ and $\rho > 1$ with possibility $1-\delta$ over the choice of $\mathcal{S}$:*

$$\mathcal{R}(s) \leq \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) + 2 \cdot \lambda_1 \cdot \mathfrak{R}_{1+2} + \sqrt{\frac{\log\log_\rho \frac{1/r_1 - 1/r_2}{1/\lambda_1 - 1/r_2}}{m}} + B \cdot \sqrt{\frac{\log(2/\delta)}{2m}},$$

$$\mathcal{R}(s) \leq \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) + 2 \cdot \lambda_1 \cdot \hat{\mathfrak{R}}_{1+2} + \sqrt{\frac{\log\log_\rho \frac{1/r_1 - 1/r_2}{1/\lambda_1 - 1/r_2}}{m}} + 3B \cdot \sqrt{\frac{\log(4/\delta)}{2m}}.$$

From existing studies, the empirical Rademacher complexity $\hat{\mathfrak{R}}_{1+2}$ often scales as $O(\sqrt{\frac{1}{m}})$. With this in mind, we can now draw the following conclusions from the theorem above. From 1), we know that the excess risk $\mathcal{R}(s) - \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s)$ is $O_p(\lambda_1 \cdot \sqrt{\frac{1}{m}})$. If we use $\tilde{f}$ instead, the corresponding result then becomes $O_p(\mathcal{LIP}(\tilde{f}) \cdot \sqrt{\frac{1}{m}})$. Consequently, using $\tilde{\ell}$ implicitly reduces the excess risk and thus improves the generalization ability. 2) states that the hyperparameter selection process will not hurt the generalization ability too much. It only increases the complexity slightly with $O(\sqrt{\log\log_\rho \frac{1/r_1 - 1/r_2}{1/\lambda_1 - 1/r_2}/m})$.

Relationship between the two frameworks. To end this subsection, we provide a discussion about the relationship between the two frameworks. As a reformulation of the first framework, the optimization problem (Re) for the contraction-loss-based framework admits the same set of the solutions with the original problem (Org). However, when deep neural networks are employed to formulate the model, the objective function is highly nonconvex. Since the optimization solution(s) for such complicated optimization problems are hardly reachable, instead of converging to the same optimal solution, these two frameworks often provide different approximations for the same optimization problem. Consequently, adopting different frameworks might end up with different performances in practical applications.



(a) Model A (b) Model B

Fig. 3: Two implementations of the proposed framework. In Model A, we implement a huber-loss-based contraction. In Model B, we implement a rlogit-loss-based-contraction.

5 Two Implementations of the Proposed Frameworks

In this section, we provide two detailed implementations for the proposed frameworks. Fig.3 gives an overview of the resulting loss functions.

5.1 Model A

5.1.1 The Sparse-Learning-Based Model

If the prior distribution of $\gamma_{i_1,i_2}^{y_{i,j}}|\lambda_1$ is a Laplacian distribution with a zero location parameter and a scale parameter of $\frac{1}{\lambda 1}$: $Lap(0, \frac{1}{\lambda_1}) = \frac{\lambda_1}{2}\exp(-\frac{|\gamma|}{1/\lambda_1})$; the prior distribution of Θ is an element-wise Gaussian distribution $\mathcal{N}(0, \frac{1}{2\lambda_2})$; and $y_{i,j}$ conditionally subjects to a Gaussian distribution $\mathcal{N}(\zeta_{i_1,i_2} + \gamma_{i_1,i_2}^{y_{i,j}}, 1)$, then the problem becomes:

$$(P_1) \min_{\Theta,\gamma} \sum_{(i,j)\in\mathcal{D}} \frac{1}{2}(\Delta_{i,j}^A - \gamma_{i_1,i_2}^{y_{i,j}})^2 + \lambda_1\|\gamma\|_1 + \lambda_2 \sum_{\theta\in\Theta} \theta^2,$$

where $\Delta^A_{i,j} = y_{i,j} - \zeta_{i_1,i_2}$.
Noteworthy is the fact that the sparse variable $\boldsymbol{\gamma}$ enjoys a clear practical meaning in this model. This could be shown formally by the following theorem, which suggests that $\boldsymbol{\gamma}$ explicitly filters out annotations that could not be represented by the learned score difference.

Theorem 4. *Denote by* $\boldsymbol{y} = [y_{i,j}]_{\{(i,j)\in\mathcal{D}\}} \in \mathbb{R}^{|\mathcal{D}|\times 1}$, $\boldsymbol{s} = [s(\boldsymbol{x}_i, \Theta)]_{\{i\in\mathcal{V}\}}$, *and* $\boldsymbol{\Psi} = [\boldsymbol{e}_{i_1,i_2}]_{\{(i,j)\in\mathcal{D}\}} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{V}|}$, *where* $\boldsymbol{e}_{i,j} \in \mathbb{R}^{1\times|\mathcal{V}|}$ *with all elements being 0 except that the* i_1*-th element being 1 and the* i_2*-th element being -1. Assume that* $rank(\boldsymbol{\Psi}) < |\mathcal{V}|$, *and the full SVD decomposition of* $\boldsymbol{\Psi}$ *is given as:*

$$\boldsymbol{\Psi} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = [\boldsymbol{U}_1, \boldsymbol{U}_2] \begin{bmatrix} \boldsymbol{\Sigma}_r, & \boldsymbol{0} \\ \boldsymbol{0}, & \boldsymbol{0} \end{bmatrix} \boldsymbol{V}^\top, \tag{4}$$

the solution of Model A could be obtained by solving the following two problems sequentially:

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \frac{1}{2}\|\mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{y}) - \mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{\gamma})\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1 \tag{5}$$

and

$$\min_{\Theta} \frac{1}{2}\|\mathsf{Proj}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}\boldsymbol{s}) - \mathsf{Proj}_{\boldsymbol{\Psi}}(\boldsymbol{y} - \hat{\boldsymbol{\gamma}})\|_2^2 + + \frac{\lambda_2}{2}\sum_{\theta\in\Theta}\theta^2, \tag{6}$$

where $\mathsf{Proj}_{\boldsymbol{\Psi}}(\cdot) = \boldsymbol{U}_1(\cdot)$ *gives a projection onto the column space of* $\boldsymbol{\Psi}$ *and the* $\mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\cdot) = \boldsymbol{U_2}(\cdot)$ *gives a projection onto the corresponding orthogonal complement.*

Remark 1. *According to Thm.4,* $\hat{\boldsymbol{\gamma}}$ *automatically detects the annotations that could not be represented as score difference with a form* $\boldsymbol{\Psi}\boldsymbol{s}$. *The HodgeRank theory [32] implies that such annotations are the sources for erroneous cyclic rankings. Such rankings include local cycles (bicycles/triangles)* $i \succ j \succ i$ / $i \succ j \succ k \succ i$ *and global cycles which involve more than three objects* $x \succ j \succ k \succ l \succ \cdots \succ i$. *With such unreliable ranking results removed, according to Eq. (6), the global ranking scores are then learned from the clean data* $\boldsymbol{y} - \hat{\boldsymbol{\gamma}}$. *The above arguments provide us the rationale to adopt this model.*

5.1.2 The Contraction-Loss-Based Model

Now we provide the loss reformulation based on Thm.1.

Corollary 1. *The* (P_1) *problem admits the same set of solutions with the following regularized Huber problem:*

$$\min_{\Theta} \sum_{(i,j)\in\mathcal{D}} \ell_{huber}(\Delta^A_{i,j}, \lambda_1) + \frac{\lambda_2}{2}\sum_{\theta\in\Theta}\theta^2, \tag{7}$$

where $\ell_{huber}(x, \lambda_1)$ *is the Huber loss defined as:*

$$\ell_{huber}(x, \lambda_1) = \begin{cases} \dfrac{x^2}{2}, & |x| \le \lambda_1 \\ \lambda_1|x| - \dfrac{\lambda_1^2}{2}, & otherwise. \end{cases} \tag{8}$$

and $\gamma^{y_{i,j}\star}_{i_1,i_2}$ *could be recovered from:*

$$\begin{cases} -\lambda_1 - \Delta^A_{i,j}, & \Delta^A_{i,j} < -\lambda_1, \\ 0, & \Delta^A_{i,j} \in [-\lambda_1, \lambda_1], \\ \Delta^A_{i,j} - \lambda_1, & \Delta^A_{i,j} > \lambda_1. \end{cases} \tag{9}$$

Proof of Corollary 1. *The proof follows Thm.1 with* $\omega_{i,j} = \Delta^A_{i,j}, \tilde{f}(t) = \frac{1}{2}t^2, \phi(t) = \frac{1}{2}t^2$. □

As shown in Fig.3a, Thm.1 reveals how robustness takes place in our model. Reformulated as a Regularized Huber problem, Model A suppresses its punishment on those unreliable examples bearing a $|\Delta^A_{i,j}|$ (the prediction deviation) larger than λ_1. This property clearly avoids the unreliable examples from dominating the whole objective function. Moreover, Eq.(9) suggests that outliers living in the cyclic ranking shown in the previous subsection could as well be found with simply comparing $\Delta^A_{i,j}$ with λ_1.

Remark 2. *Note that Huber loss, an existing loss function, is one special case in our framework. Compared with the existing findings on Huber-loss-based deep model, we further provide an efficient method to spot the outliers according to the magnitude of the loss function, the results of which are consistent with the sparse additive model. Specifically, as shown in Thm.1, the outliers in the sparse additive noise models correspond to the instances* i, j *satisfying* $\omega_{i,j} \notin [\psi(-\lambda_1), \psi(\lambda_1)]$. *More interestingly, as shown in Rem.1, with the joint effort of Hodge theory, we can show that the sparse outliers in the Huber loss can also be reformulated as the erroneous cyclic patterns in a ranking list.*

5.2 Model B

5.2.1 Sparse-Learning-Based Model

If we adopt the same assumption as above, except that we assume that $y_{i,j}$ conditionally subjects to a Logistic-like distribution, then the problem could be simplified as:

$$(P_2)\ \min_{\boldsymbol{\gamma},\Theta} \sum_{(i,j)\in\mathcal{D}} \log(1 + \Delta^B_{i,j}) + \lambda_1\|\boldsymbol{\gamma}\|_1 + \lambda_2\sum_{\theta\in\Theta}\theta^2,$$

where $\Delta^B_{i,j} = \exp(-y_{i,j}(\zeta_{i_1,i_2} + \gamma^{y_{i,j}}_{i_1,i_2}))$.

$\boldsymbol{\gamma}$ in Model B could also be interpreted with a similar spirit to Model A. Statistically, the predictive function $\frac{1}{1+\Delta^B_{i,j}}$ could be regarded as the possibility $\mathbb{P}(i_1 \overset{j}{\succ} i_2|\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2})$. Applying a logit transformation, we then reach that:

$$\log\left(\frac{\mathbb{P}(i_1 \overset{j}{\succ} i_2|\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2})}{\mathbb{P}(i_1 \overset{j}{\prec} i_2|\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2})}\right) \approxeq \Psi\boldsymbol{s} + \boldsymbol{\gamma}.$$

Here the left hand side captures the evidence to support $i_1 \overset{j}{\succ} i_2$ against $i_1 \overset{j}{\prec} i_2$. The noisy score $\Psi\boldsymbol{s} + \boldsymbol{\gamma}$ then provides a proper approximation of this

evidence. Then multiplying $\boldsymbol{U}_2$ to both sides, we then obtain:

$$\mathsf{Proj}_{\boldsymbol{\Psi}^\perp}\left(\log\left(\frac{\mathbb{P}(i_1 \overset{j}{\succ} i_2|\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2})}{\mathbb{P}(i_1 \overset{j}{\prec} i_2|\boldsymbol{x}_{i_1},\boldsymbol{x}_{i_2})}\right)\right) \approxeq \mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{\gamma}).$$

Again including the sparse noise $\boldsymbol{\gamma}$ in the model helps us to absorb the noises from the cyclic rankings.

5.2.2 Contraction-Loss-Based Model

If we restrict λ_1 in $(0,1)$, the following theorem shows that Model B also enjoys a robust reformulation.

Corollary 2. *If $\lambda_1 \in (0,1)$, the solution of (P_2) admits the same set of solutions with the following problem:*

$$\min_{\Theta} \sum_{(i,j)\in\mathcal{D}} \ell_{rlogit}(y_{i,j},\zeta_{i_1,i_2},\lambda_1) + \frac{\lambda_2}{2}\sum_{\theta\in\Theta}\theta^2, \tag{10}$$

where $\ell_{rlogit}(y_{i,j},\zeta_{i_1,i_2},\lambda_1)$ is defined as:

$$\begin{cases} \lambda_1\left|y_{i,j}\zeta_{i_1,i_2}-\xi(\lambda_1)\right| - \log(1-\lambda_1), & y_{i,j}\zeta_{i_1,i_2} < \xi(\lambda_1) \\ \log\left(1+exp(-y_{i,j}\cdot\zeta_{i_1,i_2})\right), & otherwise, \end{cases} \tag{11}$$

where $\xi(\lambda_1) = \log\left(\frac{1-\lambda_1}{\lambda_1}\right)$. Moreover, $\gamma_{i_1,i_2}^{y_{i,j}\star}$ could be recovered from:

$$\begin{cases} \xi(\lambda_1) - y_{i,j}\zeta_{i_1,i_2}, & y_{i,j}\zeta_{i_1,i_2} < \xi(\lambda_1), \\ 0, & otherwise, \end{cases} \tag{12}$$

Proof of Corollary 2. *The proof follows Thm.1-(2) with $\tilde{f}(x) = \log(1+\exp(-x)), \omega_{i,j} = y_{i,j}\zeta_{i_1,i_2}, \psi(x) = \log(-\frac{x}{1+x})$.* □

Similar to Model A, Thm.2 shows that Model B is also robust against the unreliable annotations with the construction of ℓ_{rlogit}. From the contraction loss point-of-view, the robustness is leveraged based on the notion of margin. In fact, if we regard the prediction of $y_{i,j}$ as a classification task, $y_{i,j}\zeta_{i_1,i_2}$ then serves as the well-known functional margin. As a typical property of functional margin, $y_{i,j}\zeta_{i_1,i_2}$ captures the consistency between the ground-truth and the estimated score difference ζ_{i_1,i_2}. More precisely, large positive margins correspond to correct and reliable results. By contrast, small margins (small positive value/negative value) correspond to the results that are not trustworthy. As shown in Fig.3b, ℓ_{rlogit} cleverly regards those untrustworthy annotations as "bad" samples and suppresses its punishment with a smoother linear function. Just like Model A, another merit of the reformulation comes from the implicit modeling of the parameter $\boldsymbol{\gamma}$ with only $\tilde{\ell}$ and λ_1.

6 Experiments

In this section, experiments are exhibited on three benchmark datasets (see Tab.3) which fall into two categories: (1) experiments on human age estimation from face images (Sec. 6.2), which can be considered as synthetic experiments. With the ground-truth available, this set of experiments enables us to perform in-depth evaluation of the significance of our proposed method, (2) experiments on estimating SVPs as relative attributes (Sec. 6.3 and 6.4).

TABLE 3: Dataset summary.

Dataset	No.Pairs	No.Images	No.Classes
FG-Net Face Age Dataset	15,000	1002	1
LFW-10 Dataset[28]	29,454	2000	10
Shoes Dataset [21]	87,946	14,658	7

6.1 Settings

Proposed models In the following experiments, we will evaluate our proposed **Model A** and **Model B**. Their sparse-learning-based implementations are denoted as **LS-Deep-with** $\boldsymbol{\gamma}$ and **Logit-Deep-with** $\boldsymbol{\gamma}$ respectively, while the corresponding contraction-loss-based models are named as **Huber-Deep** and **RLogit-Deep**, respectively.

Competitors We compare our method **Model A** and **Model B** with 9 stage-wise competitors.

1) **Maj-LS**: This method uses majority voting for outlier pruning and least squares problem for learning to rank.
2) **LS-with** $\boldsymbol{\gamma}$: To test the improvement of merely adopting the robust model, we jointly employ the linear regression model and our proposed robust mechanism as a baseline.
3) **Maj-Logistic**: This method stands for another baseline in our work, where the majority voting is adopted for label processing followed with the logistic regression.
4) **Logistic-with** $\boldsymbol{\gamma}$: Again, to test the improvement of merely adopting the robust model, we jointly employ the logistic regression model and our proposed robust mechanism as a baseline.
5) **Maj-RankSVM** [20]: We record the performance of RankSVM to show the superiority of the representation learning.
6) **Maj-RankNet** [4]: To show the effectiveness of using a deeper network, we compare our method with the classical RankNet model preprocessed by the majority voting.
7) **Maj-RankBoost** [8]: Besides the deep learning framework, it is also known that the ensemble-based methods could also serve a model for hierarchical learning and representation. In this sense, we compare our method with the RankBoost model, one of the most classical ensemble methods.
8) **Maj-GBDT** [9]: Gradient Boosting Decision Tree (GBDT) has gained surprising improvements in many traditional tasks. Accordingly, we compare our methods with GBDT to show its strength.
9) **Maj-DART** [31]: Recently, the well-known drop-out trick has also been applied to ensemble-based

learning, be it the DART method. We also record the performance of DART to show the superiority of our method.

For a fair comparison, we adopt the features extracted by ResNet-50 pretrained on ImageNet, which is exactly the backbone of our proposed model, in the nine traditional competitors. For the methods whose name starts with "Maj-", we first process the annotations in the training set by the majority voting and then use the results for learning.

Ablation We perform extra experiments to show whether joint end-to-end feature learning and robust ranking is better than other stage-wise deep robust ranking alternatives.

1) **pretrained+URLR** [11]: URLR is a unified robust learning to rank framework which aims to tackle both the outlier detection and learning to rank jointly. We compare our algorithm with this method to show the effectiveness of using a generalized probabilistic model and a deep architecture. In this baseline, we feed the pretrained feature extracted from ResNet-50 to URLR.
2) **noise+finetuned logit+URLR**: In this baseline, we first feed the noisy annotations to finetune a ResNet-50 network and minimize the cross-entropy loss function (logit function). After the training phase, we obtain the finetuned features from the network, which are then fed to URLR. This experiment shows us whether the noisy data is sufficient for good feature representation. Moreover, it tells us whether our proposed method outperforms finetuned features learned from noisy labels.
3) **noise+finetuned l2+URLR**: This baseline is the same as the previous one except that the loss function is changed to the squared error loss.
4) **major+finetuned logit+URLR**: In this baseline, we first perform a majority voting on the annotations and use the voted results to train a finetuned ResNet-50 network and minimize the cross-entropy loss function (logit function). After the training phase, we obtain the finetuned features from the network, which are then fed to URLR. This experiment shows us whether the majority voting procedure could remove the noises and lead to good feature representation. Moreover, it tells us whether our proposed method outperforms finetuned features learned from voted labels.
5) **major+finetuned l2+URLR**: This baseline is the same as the previous one except that the loss function is changed to the squared error loss.

Moreover, to show the effectiveness of the proposed probabilistic model, we additionally add two end-to-end competitors as the ablation. Note that the key element to detect outlier is the factor γ. In this way, the ablation competitors are formed with γ eliminated:

1) **LS-Deep-w/o γ**: This is a partial implementation of **LS-Deep-with γ**, where the factor γ is removed.
2) **Logit-Deep-w/o γ**: This is a partial implementation of **Logit-Deep-with γ**, where the factor γ is removed.

Implementation Details For each dataset, all the involved models are tuned on the validation set by grid search, and then evaluated on the test set with the best parameters. The ResNet-50 backbone in the six deep learning methods is initiated with the ImageNet pretrained parameters. For these deep methods, the learning rate is searched from $\{1e^{-5}, 1e^{-4}, 1e^{-3}\}$, and λ_2 from $\{1e^{-5}, 5e^{-4}, 1e^{-4}, 5e^{-3}, 1e^{-3}\}$. For λ_1 in LS-Deep-with-γ, we first search from $\{0.1, 0.5, 1, 1.5, 2\}$ coarsely, then make a finer tuning with a step size of 0.1 within the refined range. The coarse range for λ_1 in Logit-Deep-with-γ is $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Especially, we keep λ_1 and λ_2 for Huber-Deep (or RLogit-Deep) equal to those for LS-Deep-with-γ (or Logit-Deep-with-γ) to ensure the models' equivalence. The search ranges for other competitors are provided in appendix.

6.2 Human age dataset

In this experiment, we consider age as a subjective visual property of a face. The main difference between this SVP with the other SVPs evaluated so far is that we do have the ground-truth, i.e., the person's age when the picture was taken. This enables us to perform in-depth evaluation of the significance of our proposed framework.

Dataset The FG-NET[1] image age dataset contains 1002 images of 82 individuals labeled with ground-truth ages ranging from 0 to 69. The training set is composed of the images of 41 randomly selected individuals. For the training set, we use the ground-truth age to generate the pairwise comparisons, with the preference direction following the ground-truth order. To create sparse outliers, a random subset (i.e., 20%) of the pairwise comparisons is reversed in preference direction. In this way, we create a paired comparison graph, possibly incomplete and imbalanced, with 1002 nodes and 15,000 pairwise comparison samples. The comparisons between the remaining 41 individuals are equally divided as the validation and test set.

Evaluation metrics Because the ground-truth age is available, we adopt ACC, Precision, Recall, F1-score and AUC as the evaluation metrics to demonstrate the effectiveness of our proposed method.

Implementation Details After parameter tuning, the learning rate is set as $1e^{-5}$ for LS-Deep-w/o γ and Logit-Deep-w/o γ, $1e^{-4}$ for Logit-Deep-with γ, and $1e^{-3}$ for the rest deep models. For LS-Deep-with γ and Huber-Deep, λ_1 is set as 1.6 and λ_2 is set as $5e^{-5}$. For Logit-Deep-with γ and RLogit-Deep, λ_1 and λ_2 are set as 0.6 and $1e^{-3}$, respectively. λ_2 is set as $1e^{-4}$ for the rest deep methods.

1. http://www.fgnet.rsunit.com/

TABLE 4: Competitive results (%) on Human age dataset with the Best and Second Best results highlighted in the corresponding color.

Algorithm	Pretrain	Finetune	ACC↑	F1↑	Pre.↑	Rec.↑	AUC↑
Maj-LS	✓		72.65	65.83	64.03	67.74	79.02
LS-with γ	✓		74.54	68.13	66.39	69.96	81.42
Maj-Logistic	✓		73.46	67.00	64.86	69.27	80.41
Logistic-with γ	✓		75.28	69.13	67.20	71.18	82.68
Maj-RankNet [4]	✓		74.89	68.89	66.46	71.50	82.23
Maj-RankBoost [8]	✓		72.06	76.76	78.04	75.52	78.21
Maj-RankSVM [20]	✓		73.24	68.41	63.24	74.52	81.31
Maj-GBDT [9]	✓		74.60	66.16	68.65	63.84	81.11
Maj-DART [31]	✓		73.78	63.45	69.29	58.52	79.93
pretrained+URLR [11]	✓		76.03	69.71	68.52	70.93	83.26
noise+finetuned logit+URLR		✓	75.97	69.94	68.09	71.90	83.48
noise+finetuned l2+URLR		✓	75.47	69.14	67.67	70.67	83.28
major+finetuned logit+URLR		✓	76.06	70.07	68.20	72.04	83.52
major+finetuned l2+URLR		✓	77.32	71.55	69.86	73.33	85.30
LS-Deep-w/o γ	✓		77.21	71.22	69.98	72.50	85.10
Logit-Deep-w/o γ	✓		75.64	69.38	67.85	70.98	83.15
Huber-Deep	✓		83.80	79.34	78.69	80.00	91.89
RLogit-Deep	✓		**84.34**	**80.02**	**79.41**	80.63	92.20
LS-Deep-with γ	✓		84.31	80.00	79.30	**80.70**	**92.35**
Logit-Deep-with γ	✓		**84.57**	**80.33**	**79.63**	**81.05**	**92.50**

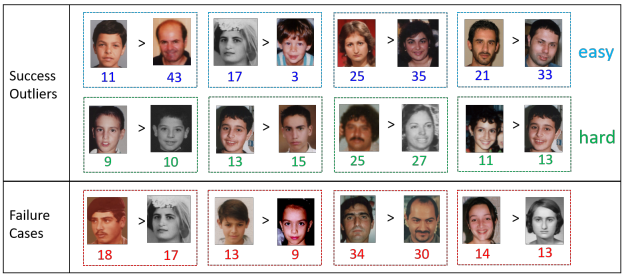
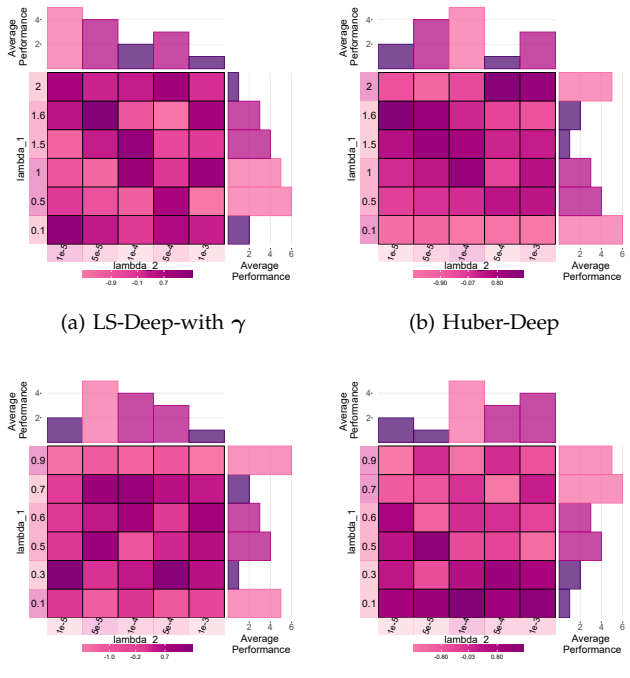


Fig. 4: Outlier examples detected on Human age dataset.

Comparative Results The competitive experiments and ablation studies are shown in Tab.4. Looking at the five-metrics results, we see that our methods consistently outperform all the benchmark algorithms by a significant margin. This validates the effectiveness of our method. In particular, it can be observed that: (1) LS-with γ (or Logistic-with γ) is superior to Maj-LS (or Maj-Logistic) because the global outlier detection is better than local outlier detection (i.e., Majority voting). (2) The finetuned feature merely gains a slight improvement with respect to the pre-trained feature. In fact, without the robust learning mechanism, the vanilla fine-tuning process (with raw/voted data) could not disentangle the contaminated patterns from the learned features. This weakens the power of traditional robust learning methods (URLR). (3) There is only a minor difference between the raw-data-based results and the majority voting-data-based results. This shows that the majority voting process fails to improve the robustness of the resulting model. As a justification, majority voting tackles the inconsistency results at a local level (removing minority directions independently). However, the higher-order/global inconsistency is totally neglected. (4) For URLR, filtering out outliers from the dataset alters the distribution of the positive/negative labeled instances. This directly results in a larger distribution gap between the training set and test set. Correspondingly, we observe a clearly worsened AUC generalization ability on the age dataset for all the five ablation studies. (5) The performance of end-to-end deep methods is better than all stage-wise methods, interestingly even the ablation baseline methods without γ give better results than traditional methods with outlier detection, which suggests that it is vital to do joint end-to-end feature learning and robust ranking. (6) Our proposed model **A** (i.e., LS-Deep-with γ and Huber-Deep) and **B** (i.e., Logit-Deep-with γ and RLogit-Deep) show comparable results on this dataset, while model **B** holds the lead by a slight margin. (7) Among the proposed models, the performance of Huber-Deep (or RLogit-Deep) is only slightly lower than LS-Deep-with γ (or Logit-Deep-with γ), which accords with the theoretical results in Sec. 4.2.

Moreover, we visualize some examples of outliers detected by LS-Deep-with γ in Fig.4, while results returned by the other proposed models are very similar. It can be seen that those in the blue/green boxes are clearly outliers and are detected correctly by our method. For better illustration, the ground-truth age is printed under each image. Moreover, blue boxes show pairs with large age differences while green boxes illustrate samples with subtle age differences, which indicates that our method not only can detect the easy pairs with a large age gap, but also can handle hard samples with small age gap (e.g., within only 1-2 years difference). Four failure cases are shown in red boxes, in which our method treats the images on the left are older than the right one as an outlier, but the ground-truth agrees with the annotation. We can easily find that this often occurs on pairs with small



(a) LS-Deep-with γ (b) Huber-Deep

(c) Logit-Deep-with γ (d) RLogit-Deep

Fig. 5: Average sensitivity analysis on Human age dataset.

age differences, which indicates that our methods may occasionally lose their power when meeting highly competitive or confusing pairs.

Sensitivity Analysis To validate the sensitivity of our model under hyperparameter perturbations, we show the results of grid search on λ_1 and λ_2 together with the best parameter for our four proposed models. The results are visualized with a 2-d heatmap and two 1-d bar plots along each parameter axis. See Fig.5a-5d. From the heatmap, we easily find that the four models have similar accuracy when using their own best parameters. The bar plot for λ_1 (λ_2) visualizes the corresponding average performance rank over all the choices of λ_2 (λ_1) with the given λ_1 (λ_2) value. We can see that for LS-Deep-with γ, $\lambda_1 = 2$ exhibits the best average performance over all the choices of λ_2, while λ_2 reaches the corresponding best performance at $1e^{-3}$. Similar observations can be found for the other three models.

Outliers Visualization From the comparative results, we have shown that Logit-Deep-with γ could achieve better performance. Here, we move a step further to explore its robustness in terms of the distribution of the magnitude of the outlier parameters, *i.e.*, $|\gamma_{i_1,i_2}^{y_{i,j}}|$. The result is shown in Fig.6. Given a pair (i_1, i_2), the x- and y- axis represent the id of image i and image i_2, respectively. For the case of convenience, we rank these image IDs in ascending order with respect to the global ranking age scores returned by our network. For each pair (i_1, i_2), let n_{i_1,i_2} be the number of comparisons, for which a_{i_1,i_2} raters agree that the age of i_2 is older than i_1 (a_{i_1,i_2} carries the opposite meaning). So $a_{i_1,i_2}+a_{i_2,i_1} = n_{i_1,i_2}$. To illustrate their distribution, we plot each a_{i_1,i_2} at (i_1, i_2) with the color and size in proportional to $|\gamma_{i_1,i_2}^{y_{i,j}}|$. It is interesting to see that the outliers picked out by our method are mainly distributed in the lower left corner $\{(i_1, i_2) : y < x\}$. The rationale behind this observation is obvious: the corresponding points indicate that $x > y$, which are in the opposite direction w.r.t the global ranking score, and thus should be considered as outliers. In this sense, we see that our proposed method could effectively pick out the reasonable outliers.

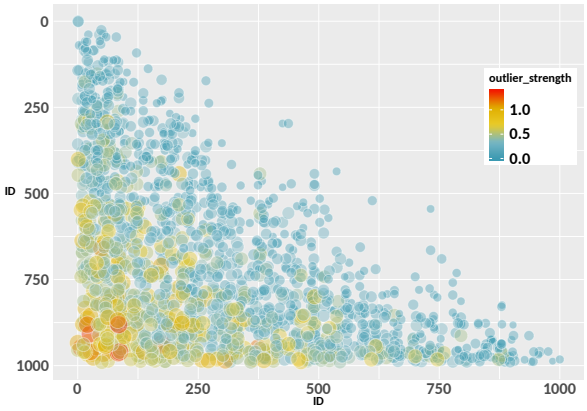


Fig. 6: Paired comparison matrix with outliers painted on Human age dataset.

Influence of Outlier Ratio To illustrate the influence of outlier ratio, we have randomly reversed 5%, 10%, 20%, 30% of generated pairwise comparisons to form different training sets for evaluation. Since the superiority of end-to-end deep learning towards the stage-wise ranking could be shown in previous experimental results, we only evaluated the six deep learning methods here. The results are visualized in Fig.7. We can see that for all the metrics, our proposed methods consistently outperform the competitors when the outlier ratio increases. This again justifies our proposed methods. Furthermore, the performance gain becomes larger when there comes more outliers, except for the ratio of 10%. The reason might be that the random 10% reverse has been applied on some very confusing pairs thus largely influenced the preference learning of the proposed models.

6.3 LFW-10 dataset

Dataset The LFW-10 dataset [28] consists of 2,000 face images, taken from the Labeled Faces in the Wild [14] dataset. It contains 10 relative attributes, like smiling, big eyes. Each pair is labeled by 5 people. For example, given a specific attribute, the user will choose which one is stronger in the attribute. As the goal of our paper is to predict SVP from noisy labels, we do not conduct any pre-processing steps to meet

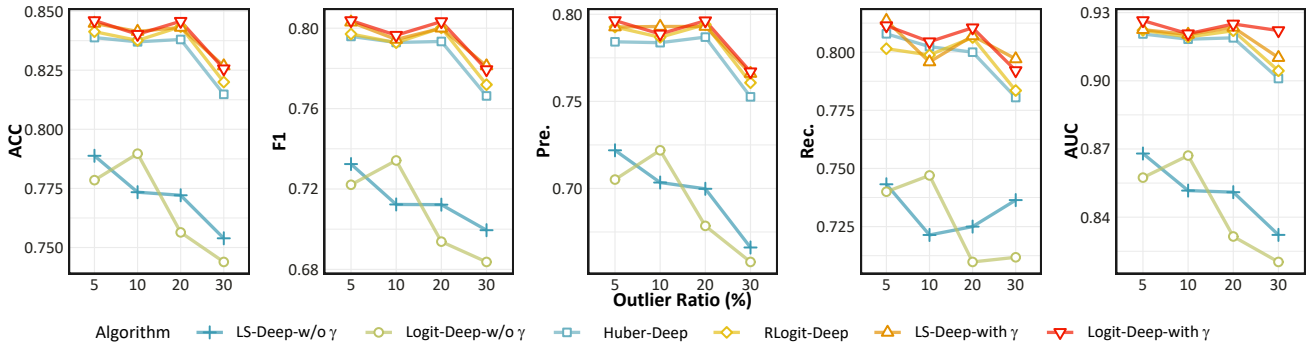


Fig. 7: Performance comparison under different outlier ratios on Human age dataset.

TABLE 5: Competitive ACC (%) results on LFW-10 dataset with the Best and Second Best results highlighted in the corresponding color.

Algorithm	Pretrain	Finetune	Bald	D.Hai	B.Eye	GLook	Masc.	Mouth	Smile	Teeth	Foreh.	Young	Aver.
Maj-LS	✓		54.62	54.58	46.61	56.49	51.82	50.64	58.16	49.79	50.66	57.43	53.08
LS-with γ	✓		50.00	59.17	43.89	62.76	56.82	54.47	58.58	54.43	59.47	62.25	56.18
Maj-Logistic	✓		52.94	62.92	44.80	63.60	58.18	52.77	59.83	56.96	58.59	66.67	57.73
Logistic-with γ	✓		49.16	63.75	43.89	64.85	61.36	54.04	62.34	59.49	63.88	64.66	58.74
Maj-RankNet [4]	✓		49.58	63.33	42.99	65.27	55.45	54.89	62.34	55.70	60.79	65.46	57.58
Maj-RankBoost [8]	✓		60.08	57.92	56.56	62.34	**67.27**	61.28	57.74	57.81	60.35	67.07	60.84
Maj-RankSVM [20]	✓		56.30	62.50	44.80	60.25	58.64	55.74	64.02	55.27	57.71	65.86	58.11
Maj-GBDT [9]	✓		59.24	57.50	55.66	61.92	**65.91**	57.87	62.34	53.59	55.07	71.89	60.10
Maj-DART [31]	✓		58.40	57.08	53.85	61.92	**65.91**	55.74	57.32	50.21	55.51	67.07	58.30
pretrained+URLR [11]	✓		49.16	59.58	53.85	61.09	65.00	57.02	61.09	58.23	62.56	63.05	59.06
noise+finetuned logit+URLR		✓	56.30	64.17	50.23	61.92	54.55	58.72	64.85	58.65	59.91	67.87	59.87
noise+finetuned l2+URLR		✓	52.52	73.75	52.04	57.74	60.00	63.40	66.11	64.56	57.27	65.86	61.45
major+finetuned logit+URLR		✓	51.26	65.83	44.80	59.00	57.73	59.15	59.83	59.92	56.83	60.24	57.57
major+finetuned l2+URLR		✓	51.68	67.92	55.20	61.09	56.82	60.00	67.36	64.14	59.47	69.48	61.45
LS-Deep-w/o γ	✓		57.98	70.83	**58.37**	59.83	58.18	65.53	66.53	64.98	62.11	69.08	63.34
Logit-Deep-w/o γ	✓		62.61	60.00	56.11	65.69	61.82	62.13	64.85	54.43	59.91	59.04	60.66
Huber-Deep	✓		59.24	74.14	57.47	59.00	61.36	**64.26**	**71.97**	65.40	**64.32**	**70.68**	**64.78**
RLogit-Deep	✓		**65.55**	67.50	51.13	**66.53**	**65.91**	59.57	63.18	60.34	59.47	64.66	62.38
LS-Deep-with γ	✓		60.08	**78.75**	**60.63**	61.92	64.09	**71.06**	67.78	**73.00**	**64.76**	**71.08**	**67.32**
Logit-Deep-with γ	✓		**65.13**	**77.08**	57.01	60.25	64.09	60.43	**68.62**	**67.09**	59.03	63.86	64.26

Fig. 8: Outlier examples of 4 representative attributes on LFW-10 dataset.

the agreement of labels as [38]. The resulting dataset has 29,454 total annotated sample pairs, on average 2945 binary pairs per attribute. For each attribute, the comparisons are randomly split by 2:1:1 into the training/validation/test set.

Implementation Details According to the grid search, the learning rate is set as $1e^{-4}$ for our LS-Deep-with γ and Logit-Deep-with γ, and $1e^{-3}$ for the rest deep models. λ_1 and λ_2 are set as 1 and $1e^{-5}$ respectively for both LS-Deep-with γ and Huber-Deep, 0.5 and $1e^{-5}$ respectively for Logit-Deep-with γ and RLogit-Deep. λ_2 is set as $1e^{-4}$ for the other two deep models.

Comparative Results Tab.5 reports the summary ACC for each attribute. The following observations can be made: (1) Our deep-methods always outperform traditional stage-wise methods and ablation baseline methods for all experiment settings with higher average ACC on all attributes. (2) Different from the results on the Human age dataset, the sparse-learning based implementation LS-Deep-with γ (or Logit-Deep-with γ) performs much better than the reformulated variant Huber-Deep (or RLogit-Deep). Recall the discussion in the last paragraph in Sec.4.2, though we have proved that LS-Deep-with γ and Huber-Deep have the same optimal solution set theoretically, the optimal solutions are hardly achievable due to the highly non-convexness of the objective function. More practically, they often provide different approximations of the solutions. Thus the perfor-

TABLE 6: Competitive ACC (%) results on Shoes dataset.

Algorithm	Pretrain	Finetune	Comf.	Fash.	Form.	Pointy	Brown	Open	Ornate	Aver.
Maj-LS	✓		81.22	84.18	79.14	**85.52**	75.47	74.26	81.68	80.21
LS-with γ	✓		80.66	83.67	78.61	**84.83**	75.00	77.21	80.63	80.09
Maj-Logistic	✓		79.56	84.18	81.28	**84.83**	76.89	75.74	80.63	80.44
Logistic-with γ	✓		79.56	83.67	81.28	**84.83**	77.83	75.00	81.15	80.47
Maj-RankNet [4]	✓		79.01	83.67	**81.82**	**84.83**	75.00	77.21	81.15	80.38
Maj-RankBoost [8]	✓		80.66	78.06	81.28	84.14	75.00	76.47	78.01	79.09
Maj-RankSVM [20]	✓		80.11	82.65	77.01	79.31	75.94	74.26	80.63	78.56
Maj-GBDT [9]	✓		80.11	79.08	**81.82**	80.69	75.47	75.00	81.68	79.12
Maj-DART [31]	✓		77.35	78.57	80.75	80.69	75.47	79.41	79.58	78.83
pretrained+URLR [11]	✓		83.94	78.82	76.89	80.00	70.05	77.52	80.77	78.39
noise+finetuned logit+URLR		✓	83.94	82.76	78.77	81.60	74.87	79.07	78.85	80.03
noise+finetuned l2+URLR		✓	**85.32**	83.74	80.19	78.40	74.87	74.42	82.21	80.42
major+finetuned logit+URLR		✓	82.11	83.74	81.60	80.80	78.61	79.07	74.52	80.11
major+finetuned l2+URLR		✓	84.86	82.76	78.30	80.80	78.61	76.74	80.77	80.66
LS-Deep-w/o γ	✓		82.87	85.71	81.28	80.69	81.60	74.26	81.68	81.16
Logit-Deep-w/o γ	✓		83.43	**87.76**	80.75	78.62	81.60	75.00	81.15	81.19
Huber-Deep	✓		81.77	86.22	**81.82**	83.45	**83.49**	**80.15**	82.20	82.73
RLogit-Deep	✓		85.08	86.22	**84.49**	79.31	83.02	75.74	**85.86**	82.82
LS-Deep-with γ	✓		83.98	**86.73**	80.75	80.69	**83.96**	**82.35**	84.29	**83.25**
Logit-Deep-with γ	✓		**85.64**	**86.73**	**81.82**	82.76	82.55	**82.35**	**84.82**	**83.81**

mance might as well turn out to be different in some cases. So do Logit-Deep-with γ and RLogit-Deep. (3) The performance of other methods is in general consistent with what we observed in the Human age experiments.

Moreover, Fig.8 gives some examples of the pruned pairs of 4 randomly selected attributes. In the success cases, the left images are (incorrectly) annotated to have more of the attribute than the right ones. However, they are either wrong or too ambiguous to give consistent answers, and as such are detrimental to learning to rank. A number of failure cases (false positive pairs identified by our models) are also shown. Some of them are caused by unique viewpoints (e.g., for 'dark hair' attribute, the man has sparse scalp, so it is hard to tell who has dark hair more); others are caused by the weak feature representation, e.g., in the 'young' attribute example, as 'young' would be a function of multiple subtle visual cues like face shape, skin texture, hair color, etc., whereas something like baldness or smiling has a better visual focus captured well by part-based features.

6.4 Shoes dataset

Dataset The Shoes dataset is collected from [21] which contains 14,658 online shopping images. In this dataset, 7 attributes are annotated by users with a wide spectrum of interests and backgrounds. For each attribute, there are at least 190 users who take part in the annotation, and each user is assigned with 50 images. Note that the dataset actually uses binary annotations rather than pairwise annotations (1 for Yes, -1 for No). We then randomly sample positive annotations and negatives annotations from each user's records to form the pairs we need. For each attribute, we randomly select such 2000 distinct pairs,

Fig. 9: Outlier examples of 4 representative attributes on Shoes dataset.

finally yielding a volume of 87,946 total personalized comparisons.

Implementation Details For all the six deep methods, the learning rate and λ_2 are set as same as those in Human age dataset, except that we set λ_2 to $1e^{-3}$ for LS-Deep-with γ and Huber-Deep, and $5e^{-5}$ for Logit-Deep-with γ and RLogit-Deep. Meanwhile, λ_1 is set as 1.2 for LS-Deep-with γ and Huber-Deep, and 0.2 for Logit-Deep-with γ and RLogit-Deep.

Comparative Results Similar to the Human age and LFW-10 datasets, Tab.6 again shows that the performance of our proposed deep models are better than that of other competitors. Moreover, some outlier detection examples are shown in Fig.9. In the top four rows with successful detection examples, the right images clearly have more of the attribute than the left ones, however are incorrectly annotated by crowdsourced raters. The failure cases are caused by the invisibility (e.g., for 'comfortable' attribute, though the transparent rain-boots itself is flat, there is in fact a pair of high-heeled shoes inside with red

color); others are caused by different visual definitions of attributes (e.g., for 'open' attribute, it has multiple shades of meaning, like peep-toed (open at toe) vs. slip-on (open at heel) vs. sandal-like (open at toe and heel)); The remaining may be caused by ambiguity: both images have this attribute with similar degree. This thus corresponds to a truly ambiguous case that can go either way.

7 Conclusion

In this paper, our goal is to mitigate the contamination issue of SVP prediction from a deep perspective. To this aim, we propose two general frameworks to simultaneously predict rank preserving scores and detect the outlier annotations. One is based on a probabilistic model which explicitly models the sparse unreliable patterns with an indicator γ, while the other reformulates the former via a "contraction" transformation of the original loss, leading to an end-to-end model with a robust loss function. For the second proposed framework, further theoretical analyses suggest its robustness and improved generalization. Then, we present two specific implementations: Model A and Model B with different assumptions on the data distribution. In our empirical studies, we perform a series of experiments on three real-world datasets: Human age dataset, LFW-10, and Shoes. The corresponding results consistently show the superiority of our proposed model.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2018AAA0102003, in part by National Natural Science Foundation of China: 61620106009, U1936208, 61733007, U1736219, 61931008, 61976202 and 61836002, in part by Key Research Program of Frontier Sciences, CAS: QYZDJ-SSW-SYS013, in part by Beijing Education Committee Cooperation Beijing Natural Science Foundation (No.KZ201910005007), in part by Youth Innovation Promotion Association CAS, and in part by the Strategic Priority Research Program of Chinese Academy of Sciences, Grant No. XDB28000000. The research of Yuan Yao was supported in part by Hong Kong Research Grant Council (HKRGC) grant 16303817, ITF UIM/390, as well as awards from Tencent AI Lab, Si Family Foundation, and Microsoft Research-Asia.

References

[1] C. G. Bampis, Z. Li, I. Katsavounidis, and A. C. Bovik. Recurrent and dynamic models for predicting streaming video quality of experience. *IEEE Transactions on Image Processing*, 27(7):3316–3331, 2018.

[2] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[3] S. Branson, G. Van Horn, and P. Perona. Lean crowdsourcing: Combining humans and machines in an online system. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7474–7483, 2017.

[4] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *International Conference on Machine Learning*, pages 89–96, 2005.

[5] K.-T. Chen, C.-C. Wu, Y.-C. Chang, and C.-L. Lei. A crowdsourceable QoE evaluation framework for multimedia content. In *ACM Conference on Multimedia*, pages 491–500, 2009.

[6] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 539–546, 2005.

[7] F. Daniel, P. Kucherbaev, C. Cappiello, B. Benatallah, and M. Allahbakhsh. Quality control in crowdsourcing: A survey of quality attributes, assessment techniques, and assurance actions. *ACM Computing Surveys*, 51(1):1–40, 2018.

[8] Y. Freund, R. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4(Nov):933–969, 2003.

[9] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[10] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao. Interestingness prediction by robust learning to rank. In *European Conference on Computer Vision*, pages 488–503, 2014.

[11] Y. Fu, T. M. Hospedales, T. Xiang, J. Xiong, S. Gong, Y. Wang, and Y. Yao. Robust subjective visual property prediction from crowdsourced pairwise labels. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):563–577, 2016.

[12] B. Han, I. W. Tsang, L. Chen, P. Y. Celina, and S.-F. Fung. Progressive stochastic learning for noisy labels. *IEEE Transactions on Neural Networks and Learning Systems*, (99):1–13, 2018.

[13] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[14] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[15] P. J. Huber. *Robust statistics*. John Wiley & Sons, 2004.

[16] X. Jiang, L.-H. Lim, Y. Yao, and Y. Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(6):203–244, 2011.

[17] Y. Jiang, Z. Yang, Q. Xu, X. Cao, and Q. Huang. When to learn what: Deep cognitive subspace clustering. In *ACM Conference on Multimedia*, pages 718–726, 2018.

[18] I. Jindal, M. Nokleby, and X. Chen. Learning deep networks from noisy labels with dropout regularization. In *IEEE International Conference on Data Mining*, pages 967–972, 2016.

[19] P. Jing, Y. Su, L. Nie, and H. Gu. Predicting image memorability through adaptive transfer learning from external sources. *IEEE Transactions on Multimedia*, 19(5):1050–1062, 2017.

[20] T. Joachims. Optimizing search engines using clickthrough data. In *ACM International Conference on Knowledge Discovery and Data Mining*, pages 133–142, 2002.

[21] A. Kovashka and K. Grauman. Discovering attribute shades of meaning with the crowd. *International Journal of Computer Vision*, 114(1):56–73, 2015.

[22] A. Kovashka and K. Grauman. Attributes for image retrieval. In *Visual Attributes*, pages 89–117. Springer, 2017.

[23] Z. Lu, Z. Fu, T. Xiang, P. Han, L. Wang, and X. Gao. Learning from weak and noisy labels for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(3):486–500, 2017.

[24] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. The MIT Press, 2018.

[25] Y. Nesterov. *Introductory lectures on convex optimization: a basic course*. Springer, 2004.

[26] M. Norouzi, D. J. Fleet, and R. R. Salakhutdinov. Hamming distance metric learning. In *Annual Conference on Neural Information Processing Systems*, pages 1061–1069, 2012.

[27] G. Patrini, A. Rozza, A. Krishna Menon, R. Nock, and L. Qu. Making deep neural networks robust to label noise: A loss

correction approach. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[28] R. N. Sandeep, Y. Verma, and C. Jawahar. Relative parts: Distinctive parts for learning relative attributes. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3614–3621, 2014.

[29] H. Squalli-Houssaini, N. Q. Duong, M. Gwenaëlle, and C.-H. Demarty. Deep learning for predicting image memorability. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2371–2375, 2018.

[30] A. Vahdat. Toward robustness against label noise in training deep discriminative neural networks. In *Annual Conference on Neural Information Processing Systems*, pages 5601–5610, 2017.

[31] R. K. Vinayak and R. Gilad-Bachrach. DART: dropouts meet multiple additive regression trees. In *International Conference on Artificial Intelligence and Statistics*, pages 489–497, 2015.

[32] Q. Xu, Q. Huang, and Y. Yao. Online crowdsourcing subjective image quality assessment. In *ACM Conference on Multimedia*, pages 359–368, 2012.

[33] Q. Xu, J. Xiong, Q. Huang, and Y. Yao. Robust evaluation for quality of experience in crowdsourcing. In *ACM Conference on Multimedia*, pages 43–52, 2013.

[34] Q. Xu, M. Yan, C. Huang, J. Xiong, Q. Huang, and Y. Yao. Exploring outliers in crowdsourced ranking for qoe. In *ACM Conference on Multimedia*, pages 1540–1548, 2017.

[35] Q. Xu, Z. Yang, Y. Jiang, X. Cao, Q. Huang, and Y. Yao. Deep robust subjective visual property prediction in crowdsourcing. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8993–9001, 2019.

[36] X. Yang, T. Zhang, C. Xu, S. Yan, M. S. Hossain, and A. Ghoneim. Deep relative attributes. *IEEE Transactions on Multimedia*, 18(9):1832–1842, 2016.

[37] J. Yao, J. Wang, I. W. Tsang, Y. Zhang, J. Sun, C. Zhang, and R. Zhang. Deep learning from noisy image labels with quality embedding. *IEEE Transactions on Image Processing*, 28(4):1909–1922, 2018.

[38] A. Yu and K. Grauman. Just noticeable differences in visual attributes. In *IEEE International Conference on Computer Vision*, pages 2416–2424, 2015.

[39] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley. Deep unsupervised saliency detection: A multiple noisy labeling perspective. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 9029–9038, 2018.

Qianqian Xu received the B.S. degree in computer science from China University of Mining and Technology in 2007 and the Ph.D. degree in computer science from University of Chinese Academy of Sciences in 2013. She is currently an Associate Professor with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. Her research interests include statistical machine learning, with applications in multimedia and computer vision. She has authored or coauthored 40+ academic papers in prestigious international journals and conferences, including T-PAMI/T-IP/T-KDE/ICML/NeurIPS/CVPR/AAAI, etc. She served as a reviewer for several top-tier journals and conferences such as T-PAMI, T-NNLS, T-MM, T-CSVT, ICML, NeurIPS, ICLR, CVPR, ECCV, AAAI, IJCAI, ACM MM, etc.

Zhiyong Yang received the M.E. degree in computer science and technology from University of Science and Technology Beijing (USTB) in 2017. He is currently pursuing the Ph.D. degree with University of Chinese Academy of Sciences. His research interests lie in theoretical and algorithmic aspects of machine learning, with special focus on multi-task learning, meta-learning, and learning with non-decomposable metrics. He has authored or coauthored several academic papers in top-tier international conferences and journals including NeurIPS/CVPR/AAAI/T-PAMI/T-IP. He served as a reviewer for several top-tier conferences such as ICML, NeurIPS and AAAI.

Yangbangyan Jiang received the bachelor's degree in instrumentation and control from Beihang University in 2017. She is currently pursuing the Ph.D. degree with University of Chinese Academy of Sciences. Her research interests include machine learning and computer vision. She has authored or coauthored several academic papers in international journals and conferences including NeurIPS, CVPR, AAAI, etc. She served as a reviewer for several top-tier conferences such as ICML, NeurIPS, ICLR, CVPR, AAAI.

Xiaochun Cao, Professor of the Institute of Information Engineering, Chinese Academy of Sciences. He received the B.E. and M.E. degrees both in computer science from Beihang University (BUAA), China, and the Ph.D. degree in computer science from the University of Central Florida, USA, with his dissertation nominated for the university level Outstanding Dissertation Award. After graduation, he spent about three years at ObjectVideo Inc. as a Research Scientist. From 2008 to 2012, he was a professor at Tianjin University. He has authored and coauthored over 100 journal and conference papers. In 2004 and 2010, he was the recipients of the Piero Zamperoni best student paper award at the International Conference on Pattern Recognition. He is a fellow of IET and a Senior Member of IEEE. He is an associate editor of IEEE Transactions on Image Processing, IEEE Transactions on Circuits and Systems for Video Technology and IEEE Transactions on Multimedia.

Yuan Yao received the B.S.E and M.S.E in control engineering both from Harbin Institute of Technology, China, in 1996 and 1998, respectively, M.Phil in mathematics from City University of Hong Kong in 2002, and Ph.D. in mathematics from the University of California, Berkeley, in 2006. Since then he has been with Stanford University and in 2009, he joined the Department of Probability and Statistics in School of Mathematical Sciences, Peking University, Beijing, China. He is currently an Associate Professor of Mathematics, Chemical & Biological Engineering, and by courtesy, Computer Science & Engineering, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong SAR, China. His current research interests include topological and geometric methods for high dimensional data analysis and statistical machine learning, with applications in computational biology, computer vision, and information retrieval. Dr. Yao is a member of American Mathematical Society (AMS), Association for Computing Machinery (ACM), Institute of Mathematical Statistics (IMS), and Society for Industrial and Applied Mathematics (SIAM). He served as area or session chair in NIPS and ICIAM, as well as a reviewer of Foundation of Computational Mathematics, IEEE Trans. Information Theory, J. Machine Learning Research, and Neural Computation, etc.

Qingming Huang is a chair professor in University of Chinese Academy of Sciences and an adjunct research professor in the Institute of Computing Technology, Chinese Academy of Sciences. He graduated with a Bachelor degree in Computer Science in 1988 and Ph.D. degree in Computer Engineering in 1994, both from Harbin Institute of Technology, China. His research areas include multimedia computing, image processing, computer vision and pattern recognition. He has authored or coauthored more than 400 academic papers in prestigious international journals and top-level international conferences. He is the associate editor of IEEE Trans. on CSVT and Acta Automatica Sinica, and the reviewer of various international journals including IEEE Trans. on PAMI, IEEE Trans. on Image Processing, IEEE Trans. on Multimedia, etc. He is a Fellow of IEEE and has served as general chair, program chair, track chair and TPC member for various conferences, including ACM Multimedia, CVPR, ICCV, ICME, ICMR, PCM, BigMM, PSIVT, etc.

APPENDIX

OPTIMIZATION DETAILS IN SEC.4

Now, we elaborate how the first framework for Model A and Model B could be trained in an alternating optimization strategy.

Model A

Fix γ, Learn Θ. When fixing γ, we see that Θ could be solved from the following subproblem:

$$\min_{\Theta} -\sum_{(i,j)\in\mathcal{D}} \log(f_{ij}) - \log(g)$$

Since Θ only depends on the network, one could find an approximated solution by off-the-shelf deep learning tools. For **Model A**, this subproblem becomes:

$$\min_{\Theta} \sum_{(i,j)\in\mathcal{D}} \frac{1}{2}\left(\Delta_{i,j}^{A} - \gamma_{i_1,i_2}^{y_{i,j}}\right)^2 + \lambda_2 \sum_{\theta\in\Theta} \theta^2.$$

Fix Θ, Learn γ. Similarly, when Θ is fixed, we could solve γ from:

$$\min_{\gamma} \sum_{(i,j)\in\mathcal{D}} -\left(\log(f_{ij}) + \log(h_{ij})\right)$$

This is a simple model of γ which is not entangled with the network. For **Model A**, this subproblem becomes:

$$\min_{\gamma} \sum_{(i,j)\in\mathcal{D}} \frac{1}{2}\left(\Delta_{i,j}^{A} - \gamma_{i_1,i_2}^{y_{i,j}}\right)^2 + \lambda_1\|\gamma\|_1.$$

It enjoys a closed-form solution with the proximal operator of ℓ_1 norm:

$$\gamma_{i_1,i_2}^{y_{i,j}} = \max\left(|\Delta_{i,j}^{A}| - \lambda_1, 0\right)\cdot \operatorname{sign}\left(\Delta_{i,j}^{A}\right). \tag{13}$$

Model B

Fix γ, Learn Θ. Similarly, for **Model B**, this results in the following subproblem:

$$\min_{\Theta} \sum_{(i,j)\in\mathcal{D}} \log(1+\Delta_{i,j}^{B}) + \lambda_2 \sum_{\theta\in\Theta} \theta^2.$$

which could be solved by a standard optimizer for deep learning.

Fix Θ, Learn γ. For **Model B**, this γ subproblem becomes:

$$\min_{\gamma} \sum_{(i,j)\in\mathcal{D}} \log(1+\Delta_{i,j}^{B}) + \lambda_1\|\gamma\|_1,$$

Unfortunately, there is no closed-form solution for this subproblem. In this paper, we adopt the proximal gradient method [2] to find a numerical solution. As a preliminary, let us first show that the objective function has $\frac{1}{4}$-Lipschitz continuous gradient toward γ.

Proposition 1. *Denote ℓ_{logit} by $\min_{\Theta}\sum_{(i,j)\in\mathcal{D}}\log(1+\Delta_{i,j}^{B})$, then $\nabla_{\gamma}\ell_{logit}$ is $\frac{1}{4}n_{i_1,i_2 max}^{y_{i,j}}$-Lipschitz continuous, where $n_{i_1,i_2 max}^{y_{i,j}} = \max_{i_1,i_2,y_{i,j}} {n_{i_1,i_2}^{y_{i,j}}}^2$.*

Proof of Proposition 1. *Taking the second-order derivative with respect to $\gamma_{i_1,i_2}^{y_{i,j}}$, we have*

$$\frac{\partial^2\ell_{logit}}{\partial {\gamma_{i_1,i_2}^{y_{i,j}}}^2} = \left(n_{i_1,i_2}^{y_{i,j}}\right)^2 \cdot sig_{ij}\cdot(1-sig_{ij})$$

where $sig_{ij} = \dfrac{1}{1+\Delta_{i,j}^{B}}$. Moreover, since $0 \le sig_{ij} \le 1$, we have $\dfrac{\partial^2\ell_{logit}}{\partial {\gamma_{i_1,i_2}^{y_{i,j}}}^2} \le \dfrac{1}{4}n_{i_1,i_2 max}^{y_{i,j}}$. Denote $\nabla_{\gamma}^2\ell_{logit}$ as the corresponding Hessian matrix, we have

$$\nabla_{\gamma}^2\ell_{logit} = diag\left(\left[\frac{\partial\ell_{logit}}{\partial\gamma_{i_1,i_2}^{y_{i,j}}}\right]_{\{(i,j)\}\in\mathcal{D}}\right) \preceq \frac{n_{i_1,i_2 max}^{y_{i,j}}\boldsymbol{I}}{4}.$$

This implies that $||\nabla_{\gamma}^2\ell_{logit}||_2 \le \dfrac{n_{i_1,i_2 max}^{y_{i,j}}}{4}$. According to Lem. 1.2.2 of [25], we know that $\nabla_{\gamma}\ell_{logit}$ is $\frac{1}{4}n_{i_1,i_2 max}^{y_{i,j}}$-Lipschitz continuous. □

According to the proximal gradient descent method, the solution of this subproblem could be obtained by iteratively solving :

$$\boldsymbol{\gamma}^{(t)} = \operatorname*{argmin}_{\gamma}\frac{1}{2}||\boldsymbol{\gamma} - \widetilde{\boldsymbol{\gamma}}^{(t-1)}||_2^2 + \lambda_1||\boldsymbol{\gamma}||_1, \tag{14}$$

where $\widetilde{\boldsymbol{\gamma}}^{(t-1)} = \boldsymbol{\gamma}^{(t-1)} - \dfrac{4}{n_{i_1,i_2 max}^{y_{i,j}}}\cdot\nabla_{\boldsymbol{\gamma}}\ell_{logit}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}^{(t-1)}}$.

PROOFS

Restate of Theorem 1 (Unified Reformulation of the Sparse Additive Noise Model). *Considering the model formulation in Eq.(2), we set $h_{ij} = \exp(-\lambda_1\cdot|\gamma_{i_1,i_2}^{y_{i,j}}|)$, $g = \prod_{\theta\in\Theta}\exp(\lambda_2\cdot\theta^2)$ and $\log(f_{i,j}) = \tilde{f}\left(\omega(y_{i,j},\zeta_{i_1,i_2}) + \gamma_{i_1,i_2}^{y_{i,j}}\right)$, where $\tilde{f}$ is a strictly convex function such that $\tilde{f}' = \frac{d\tilde{f}(x)}{dx}$ is invertible, continuous and strictly increasing, and $\mathsf{Range}(\omega) = \mathbb{R}$. For all λ_1 such that at least one element in $\{-\lambda_1,\lambda_1\}$ belongs to $\mathsf{Range}(\tilde{f})$, the original problem shares the same solution set with the following problem (Re):*

$$(Re)\quad \operatorname*{argmin}_{\Theta}\sum_{(i,j)\in\mathcal{D}}\tilde{\ell}(\omega_{i,j},\lambda_1) + \lambda_2\cdot\sum_{\theta\in\Theta}\theta^2.$$

If we further define $\psi(x) = \tilde{f}'^{-1}(x), \phi(\cdot) = \tilde{f}\left(\psi(\cdot)\right)$ and $\omega_{i,j} = \omega\left(y_{i,j},\zeta_{i_1,i_2}\right)$, then the following facts hold:

(a) $\tilde{\ell}(\omega_{i,j},\lambda_1)$ could be expressed as:

(1) If $[-\lambda,\lambda]\subseteq\mathsf{Range}(\tilde{f}')$,

$$\begin{cases}\phi(-\lambda_1)+\lambda_1\cdot|\omega_{i,j}-\psi(-\lambda_1)|, & \omega_{i,j}<\psi(-\lambda_1),\\ \phi(\lambda_1)+\lambda_1\cdot|\omega_{i,j}-\psi(\lambda_1)|, & \omega_{i,j}>\psi(\lambda_1),\\ \tilde{f}(\omega_{i,j}), & otherwise,\end{cases}$$

(2) If $-\lambda\notin\mathsf{Range}(\tilde{f}')$,

$$\begin{cases}\tilde{f}(\omega_{i,j}), & \omega_{i,j}\le\psi(\lambda_1),\\ \phi(\lambda_1)+\lambda_1\cdot|\omega_{i,j}-\psi(\lambda_1)|, & \omega_{i,j}>\psi(\lambda_1).\end{cases}$$

(3) If $\lambda \notin \text{Range}(\tilde{f}')$,

$$\begin{cases} \phi(-\lambda_1) + \lambda_1 \cdot |\omega_{i,j} - \psi(-\lambda_1)|, & \omega_{i,j} < \psi(-\lambda_1), \\ \tilde{f}(\omega_{i,j}), & \omega_{i,j} \geq \psi(-\lambda_1), \end{cases}$$

(b) Given any feasible Θ, the partial optimal solution $\gamma_{i_1,i_2}^{y_{i,j}\star}$ is:

(1) If $[-\lambda, \lambda] \subseteq \text{Range}(\tilde{f}')$,

$$\begin{cases} \psi(-\lambda_1) - \omega_{i,j}, & \omega_{i,j} < \psi(-\lambda_1), \\ 0, & \omega_{i,j} \in [\psi(-\lambda_1), \psi(\lambda_1)], \\ \omega_{i,j} - \psi(\lambda_1), & \omega_{i,j} > \psi(\lambda_1). \end{cases}$$

(2) If $-\lambda \notin \text{Range}(\tilde{f}')$,

$$\begin{cases} 0, & \omega_{i,j} \leq \psi(\lambda_1), \\ \omega_{i,j} - \psi(\lambda_1), & \omega_{i,j} > \psi(\lambda_1). \end{cases}$$

(3) If $\lambda \notin \text{Range}(\tilde{f}')$,

$$\begin{cases} \psi(-\lambda_1) - \omega_{i,j}, & \omega_{i,j} < \psi(-\lambda_1), \\ 0, & \omega_{i,j} \geq \psi(-\lambda_1). \end{cases}$$

Proof of Theorem 1. *First, reformulate (Re) as a bilevel optimization problem:*

$$\min_{\Theta} \sum_{(i,j)\in\mathcal{D}} \cdot \Big(\min_{\substack{y_{i,j} \\ \gamma_{i_1,i_2}}} \tilde{f}(\omega\,(y_{i,j}, \zeta_{i_1,i_2}) + \gamma_{i_1,i_2}^{y_{i,j}}) + \lambda_1 |\gamma_{i_1,i_2}^{y_{i,j}}|_1 \Big) + \lambda_2 \cdot \sum_{\theta\in\Theta} \theta^2.$$

Now, we only need to focus on the inner problem minimizing the individual $\gamma_{i_1,i_2}^{y_{i,j}}$s. Since $\tilde{f}$ is, by assumption, strictly convex, the optimal $\gamma_{i_1,i_2}^{y_{i,j}}$ must satisfy:

$$0 \in \tilde{f}'(\omega\,(y_{i,j}, \zeta_{i_1,i_2}) + \gamma_{i_1,i_2}^{y_{i,j}}) + \lambda_1 \cdot \partial(|\gamma_{i_1,i_2}^{y_{i,j}}|), \quad (15)$$

where the subdifferential of $|\gamma_{i_1,i_2}^{y_{i,j}}|$ is defined as:

$$\partial(|\gamma_{i_1,i_2}^{y_{i,j}}|) = \begin{cases} 1, & \gamma_{i_1,i_2}^{y_{i,j}} > 0, \\ [-1, 1], & \gamma_{i_1,i_2}^{y_{i,j}} = 0, \\ -1, & \gamma_{i_1,i_2}^{y_{i,j}} < 0. \end{cases}$$

It is easy to see that the function $\tilde{\ell}$ now could be derived from the minimum value of the inner problem.
Proof of (a):
First, we prove (1), which could be finished with the following three cases:
Case 1 *$\gamma_{i_1,i_2}^{y_{i,j}} > 0$. Eq.(15) then suggests that $\tilde{f}'(\omega_{i,j} + \gamma_{i_1,i_2}^{y_{i,j}}) = -\lambda_1$, which means $\gamma_{i_1,i_2}^{y_{i,j}} = \tilde{f}'^{-1}(-\lambda_1) - \omega_{i,j}$. Moreover, $\gamma_{i_1,i_2}^{y_{i,j}} > 0$ implies that $\omega_{i,j} < \tilde{f}'^{-1}(-\lambda_1)$. By substituting the expression of $\gamma_{i_1,i_2}^{y_{i,j}}$ into the loss function, we can reach the corresponding result.*
Case 2 *$\gamma_{i_1,i_2}^{y_{i,j}} = 0$. Eq.(15) then suggests that $-\lambda_1 \leq \tilde{f}'(\omega_{i,j} + \gamma_{i_1,i_2}^{y_{i,j}}) \leq \lambda_1$, which means $\omega_{i,j} \in [\tilde{f}'^{-1}(-\lambda_1), \tilde{f}'^{-1}(\lambda_1)]$. Again, $\gamma_{i_1,i_2}^{y_{i,j}} = 0$ leads to the corresponding result.*
Case 3 *$\gamma_{i_1,i_2}^{y_{i,j}} < 0$. The proof is similar to **Case 1**.*
For (2), we only need to realize that since $\tilde{f}'$ is continuous, $\lambda_1 \in \text{Range}(\tilde{f}')$ and $-\lambda_1 \notin \text{Range}(\tilde{f}')$ implies that $\forall x \in \text{Range}(\tilde{f}'), x > -\lambda_1$. In other words, Case 3 in (1) is not achievable and $-\lambda_1 \leq \tilde{f}'(\omega_{i,j} + \gamma_{i_1,i_2}^{y_{i,j}})$ holds for all $\omega_{i,j}$. Applying these two facts to the proof of (1), we can then complete the proof.
The proof of (3) follows the same spirit of (2) and thus is omitted here.
Proof of (b) is finished in the proof of (a). □

Restate of Theorem 2 (Robustness of the Loss function). *Under the assumptions of Thm.1, $\tilde{\ell}$ has the following properties:*

(1) For all

$$\omega_{i,j} \in \mathcal{H} = \{\omega_{i,j} : \omega_{i,j} < \psi(-\lambda_1) \text{ or } \omega_{i,j} > \psi(\lambda_1)\},$$

we have:

$$\tilde{\ell}(\omega_{i,j}, \lambda_1) < \tilde{f}(\omega_{i,j})$$

(2) $|\tilde{\ell}(\omega_{i,j}, \lambda_1) - \tilde{\ell}(\omega'_{i,j}, \lambda_1)| \leq |\tilde{f}(\omega_{i,j}) - \tilde{f}(\omega'_{i,j})|$, with inequality holds strictly if at least element in $\{\omega_{i,j}, \omega'_{i,j}\}$ belongs to $\mathcal{H}$.
(3) $\mathcal{LIP}(\tilde{\ell}(\cdot, \lambda_1)) = \lambda_1$, $\mathcal{LIP}(\tilde{f}) > \lambda_1$.

Proof of Theorem 2.
(1)*: Denote $\mathcal{H}_1$ by $\{\omega_{i,j} : \omega_{i,j} < \psi(-\lambda_1)\}$ and denote $\mathcal{H}_2$ by $\{\omega_{i,j} : \omega_{i,j} > \psi(\lambda_1)\}$. If $\omega_{i,j} \in \mathcal{H}_1$, we have $\tilde{f}'(\omega_{i,j}) < -\lambda_1$. Moreover, since $\tilde{f}'$ is strictly increasing, we have $\tilde{f}'(\omega_{i,j}) < \tilde{f}'(\omega'_{i,j})$, if $\omega'_{i,j} < \omega_{i,j}$. By choosing $\omega'_{i,j}$ as $\psi(-\lambda_1)$, we have that $\tilde{f}'(\omega_{i,j}) < -\lambda_1$, if $\omega_{i,j} < \psi(-\lambda_1)$. According to the definition of $\tilde{\ell}$, we have $\tilde{f}(\omega'_{i,j}) = \tilde{\ell}(\omega'_{i,j}, \lambda_1) = \phi(-\lambda_1)$ when $\omega'_{i,j} = \psi(-\lambda_1)$. Putting all together, we have $\forall \omega_{i,j} \in \mathcal{H}_1$,*

$$\begin{aligned} \tilde{f}(\omega_{i,j}) &= \tilde{f}(\omega_{i,j}) - \tilde{f}(\omega'_{i,j}) + \tilde{f}(\omega'_{i,j}) \\ &= \tilde{f}(\omega_{i,j}) - \phi(-\lambda_1) + \phi(-\lambda_1) \\ &= \phi(-\lambda_1) - \int_{\omega_{i,j}}^{\psi(-\lambda_1)} \tilde{f}'(t)dt \\ &> \phi(-\lambda_1) + \int_{\omega_{i,j}}^{\psi(-\lambda_1)} \lambda_1 dt \\ &= \phi(-\lambda_1) + \lambda_1 \cdot (\psi(-\lambda_1) - \omega_{i,j}) \\ &= \tilde{\ell}(\omega_{i,j}, \lambda_1). \end{aligned} \quad (16)$$

The proof for $\omega_{i,j} \in \mathcal{H}_2$ could be finished with a similar spirit.
(2)*: This is a direct result from (1).*
(3)*: $\mathcal{LIP}(\tilde{\ell}(,\lambda_1)) = \lambda_1$ follows the fact that $-\lambda_1 \leq \tilde{\ell}'(\omega_{i,j}, \lambda_1) \leq \lambda_1$, when $\omega_{i,j} \notin \mathcal{H}$ and $|\tilde{\ell}'(\omega_{i,j}, \lambda_1)| = \lambda_1$, $\omega_{i,j} \in \mathcal{H}$. $\mathcal{LIP}(\tilde{f}) > \lambda_1$ follows from the fact that $|\tilde{f}'(\omega_{i,j})| > \lambda_1, \forall \omega_{i,j} \in \mathcal{H}$.* □

Restate of Theorem 3 (Improved Generalization from Robustness). *Let $\mathcal{F}$ be the hypothesis space for the real scoring function $s(\cdot)$, such that $\zeta_{i_1,i_2} = s(\boldsymbol{x}_{i_1}) - s(\boldsymbol{x}_{i_2})$. Define the i.i.d training sample as $\mathcal{S} = \{(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}, y_{i,j})\}_{i_1,i_2,j}$, and define two partial sets as $\mathcal{S}_1 = \{(\boldsymbol{x}_{i_1}, y_{i,j})\}_{i_1,j}$, and $\mathcal{S}_2 = \{(\boldsymbol{x}_{i_2}, y_{i,j})\}_{i_2,j}$. Moreover, let*

$$\mathfrak{R}_{\mathcal{S}_1}(\mathcal{F}) = \mathbb{E}_{\mathcal{S}}(\hat{\mathfrak{R}}_{\mathcal{S}_1}(\mathcal{F})),\ \mathfrak{R}_{\mathcal{S}_2}(\mathcal{F}) = \mathbb{E}_{\mathcal{S}}(\hat{\mathfrak{R}}_{\mathcal{S}_2}(\mathcal{F})),$$
$$\mathfrak{R}_{1+2} = \mathfrak{R}_{\mathcal{S}_1}(\mathcal{F}) + \mathfrak{R}_{\mathcal{S}_2}(\mathcal{F}),\ \hat{\mathfrak{R}}_{1+2} = \hat{\mathfrak{R}}_{\mathcal{S}_1}(\mathcal{F}) + \hat{\mathfrak{R}}_{\mathcal{S}_2}(\mathcal{F})$$
$$\hat{\mathcal{R}}_{\mathcal{S},\lambda} = \tfrac{1}{m} \sum_{(i,j)\in\mathcal{D}} \tilde{\ell}(\omega_{i,j}, \lambda),\ \mathcal{R}(s) = \mathbb{E}_{\mathcal{S}}[\sum_{(i,j)\in\mathcal{D}} \mathbf{1}_{y_{i,j}\cdot(\zeta_{i_1,i_2}) \leq 0}]$$

The following two results about the generalization ability for $\tilde{\ell}$ based ERM hold:

1) *Under the assumption of Thm.1 if $\omega_{i,j} = y_{i,j} \cdot \zeta_{i_1,i_2}$ and $\mathbf{1}_{x\leq 0} \leq \tilde{\ell}(x,\lambda_1)$, $\mathsf{Range}(\tilde{\ell}) = [0, B]$, with a fixed $\lambda_1 > 0$, for any $\delta \in (0,1)$, the following facts hold for all $s \in \mathcal{F}$ with probability at least $1-\delta$ over the choice of the training sample $\mathcal{S}$:*

$$\mathcal{R}(s) \leq \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) + 2\cdot\lambda_1\cdot\mathfrak{R}_{1+2} + B\cdot\sqrt{\frac{\log(1/\delta)}{2m}}$$

$$\mathcal{R}(s) \leq \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) + 2\cdot\lambda_1\cdot\hat{\mathfrak{R}}_{1+2} + 3B\cdot\sqrt{\frac{\log(2/\delta)}{2m}}$$

2) *Under the same assumption in (1), $\forall\delta \in (0,1)$ the following inequalities hold for all $\lambda_1 \in [r_1, r_2)$ and $\rho > 1$ with possibility $1-\delta$ over the choice of $\mathcal{S}$:*

$$\mathcal{R}(s) \leq \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) + 2\cdot\lambda_1\cdot\mathfrak{R}_{1+2} + \sqrt{\frac{\log\log_\rho \frac{1/r_1-1/r_2}{1/\lambda_1-1/r_2}}{m}} + B\cdot\sqrt{\frac{\log(2/\delta)}{2m}},$$

$$\mathcal{R}(s) \leq \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) + 2\cdot\lambda_1\cdot\hat{\mathfrak{R}}_{1+2} + \sqrt{\frac{\log\log_\rho \frac{1/r_1-1/r_2}{1/\lambda_1-1/r_2}}{m}} + 3B\cdot\sqrt{\frac{\log(4/\delta)}{2m}}.$$

Proof of Theorem 3.
(1). *This follows the Eq.(3) and [24, Thm.10.1].*
(2). *Assume that we have two sequences $\{\lambda_{1,k}\}_{k\in\mathbb{N}_+}$, $\{\epsilon_k\}_{k\in\mathbb{N}_+}$, with $\epsilon_k = \epsilon + \sqrt{\frac{\log k}{m}}$. From the result of (1), we have:*

$$\begin{aligned}
&\mathbb{P}\left[\sup_{s\in\mathcal{F},k\geq 1} \mathcal{R}(s) - \hat{\mathcal{R}}_{\mathcal{S},\lambda_{1,k}}(s) - \lambda_{1,k}\mathfrak{R}_{1+2} - \epsilon_k \geq 0\right] \\
&\leq \sum_{k\geq 1}\exp(-2m\epsilon_k^2) \\
&= \sum_{k\geq 1}\exp\left(-2m\left(\epsilon + \sqrt{\frac{\log k}{m}}\right)\right) \\
&\leq \exp(-2m\epsilon^2)\cdot\left(\sum_{k\geq 1}\exp(-2\log k)\right) \\
&\leq 2\exp(-2m\epsilon)
\end{aligned}$$

Now given any $\lambda_1 \in [r_1, r_2)$, we now continue to find a corresponding $\epsilon_k, \lambda_{1,k}$ in the sequence. Now we construct the sequence of $\lambda_{1,k}$ as $1/\lambda_{1,k} = 1/r_2 + \frac{1/r_1-1/r_2}{\rho^k}$. Then, for any $\lambda_1 \in [r_1, r_2)$, there exists a k such that $\lambda_1 \in [\lambda_{1,k}, \lambda_{1,k+1}]$. Here $\tilde{\ell}(\cdot,\lambda_1) > \tilde{\ell}(\cdot,\lambda_{1,k})$. Since $1/\lambda_{1,k} > 1/\lambda_1$, we have $k = \log_\rho \frac{1/r_1-1/r_2}{1/\lambda_{1,k}-1/r_1} \leq \log_\rho \frac{1/r_1-1/r_2}{1/\lambda_1-1/r_2}$. All the above results hold consistently for any $\lambda_1 \in [r_1, r_2)$, we then reach the inequality:

$$\begin{aligned}
&\mathbb{P}\Bigg[\sup_{\substack{s\in\mathcal{F}\\ \lambda_1\in[r_1,r2)}} \mathcal{R}(s) - \hat{\mathcal{R}}_{\mathcal{S},\lambda_1}(s) - \lambda_1\mathfrak{R}_{1+2} - \epsilon - \\
&\sqrt{\frac{\log\log_\rho \frac{1/r_1-1/r_2}{1/\lambda_1-1/r_2}}{m}} \geq 0\Bigg] \\
&\leq \mathbb{P}\left[\sup_{\substack{s\in\mathcal{F}\\ k\geq 1}} \mathcal{R}(s) - \hat{\mathcal{R}}_{\mathcal{S},\lambda_{1,k}}(s) - \lambda_{1,k}\mathfrak{R}_{1+2} - \epsilon_k \geq 0\right] \\
&\leq 2\exp(-2m\epsilon^2)
\end{aligned}$$

This ends the proof of the first inequality of (2). The second inequality simply follows a concentration over the Rademacher complexity. □

Restate of Theorem 4. *Denote by $\boldsymbol{y} = [y_{i,j}]_{\{(i,j)\in\mathcal{D}\}} \in \mathbb{R}^{|\mathcal{D}|\times 1}$, $\boldsymbol{s} = [s(\boldsymbol{x}_i,\Theta)]_{\{i\in\mathcal{V}\}}$, and $\boldsymbol{\Psi} = [\boldsymbol{e}_{i_1,i_2}]_{\{(i,j)\in\mathcal{D}\}} \in \mathbb{R}^{|\mathcal{D}|\times|\mathcal{V}|}$, where $\boldsymbol{e}_{i,j} \in \mathbb{R}^{1\times|\mathcal{V}|}$ with all elements being 0 except that the i_1-th element being 1 and the i_2-th element being -1. Assume that $rank(\boldsymbol{\Psi}) < |\mathcal{V}|$, and the full SVD decomposition of $\boldsymbol{\Psi}$ is given as:*

$$\boldsymbol{\Psi} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = [\boldsymbol{U}_1, \boldsymbol{U}_2]\begin{bmatrix}\boldsymbol{\Sigma}_r, & \mathbf{0}\\ \mathbf{0}, & \mathbf{0}\end{bmatrix}\boldsymbol{V}^\top, \tag{17}$$

the solution of Model A could be obtained by solving the following two problems sequentially:

$$\hat{\boldsymbol{\gamma}} = \underset{\boldsymbol{\gamma}}{\operatorname{argmin}} \frac{1}{2}\|\mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{y}) - \mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{\gamma})\|_2^2 + \lambda_1\|\boldsymbol{\gamma}\|_1 \tag{18}$$

and

$$\min_{\Theta} \frac{1}{2}\|\mathsf{Proj}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}\boldsymbol{s}) - \mathsf{Proj}_{\boldsymbol{\Psi}}(\boldsymbol{y}-\hat{\boldsymbol{\gamma}})\|_2^2 + \frac{\lambda_2}{2}\sum_{\theta\in\Theta}\theta^2, \tag{19}$$

where $\mathsf{Proj}_{\boldsymbol{\Psi}}(\cdot) = \boldsymbol{U}_1(\cdot)$ gives a projection onto the column space of $\boldsymbol{\Psi}$ and the $\mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\cdot) = \boldsymbol{U}_2(\cdot)$ gives a projection onto the corresponding orthogonal complement.

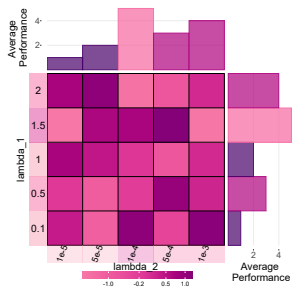
Proof of Theorem 4. *One can easily rearrange the objective function of Model A as:*

$$\min_{\Theta,\boldsymbol{\gamma}} \frac{1}{2}||\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{s} - \boldsymbol{\gamma}||_2^2 + \lambda_1\|\boldsymbol{\gamma}\|_1 + \frac{\lambda_2}{2}\sum_{\theta\in\Theta}\theta^2. \tag{20}$$

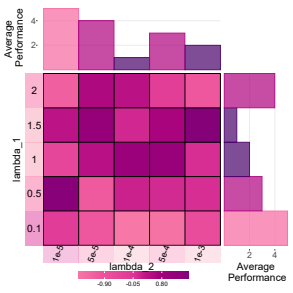
Since $span([\boldsymbol{U}_1, \boldsymbol{U}_2]) = \mathbb{R}^{|\mathcal{D}|}$, for any $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{D}|}$, we have $\boldsymbol{z} = \boldsymbol{U}_1(\boldsymbol{z}_1) + \boldsymbol{U}_2(\boldsymbol{z}_2)$. Then we know that $\mathsf{Proj}_{\boldsymbol{\Psi}}(\boldsymbol{z}) = \boldsymbol{z}_1$ and $\mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{z}) = \boldsymbol{z}_2$ are the coordinates of $\boldsymbol{z}$ for basis $\boldsymbol{U}_1$ and $\boldsymbol{U}_2$, respectively. Practically, according to the fundamental properties of SVD decomposition, $\boldsymbol{U}_1$ spans the column space of $\boldsymbol{\Psi}$ while $\boldsymbol{U}_2$ spans the corresponding orthogonal complement. According to the unitary invariance of the ℓ_2 vector norm, we have $||\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{s} - \boldsymbol{\gamma}||_2^2 = ||\mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{y}) - \mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{\gamma})||_2^2$ by the fact that $\mathsf{Proj}_{\boldsymbol{\Psi}^\perp}(\boldsymbol{\Psi}\boldsymbol{s}) = \mathbf{0}$. This shows that $\boldsymbol{\gamma}$ could be solved from Eq. (18). Moreover, by plugging in the resulting $\hat{\boldsymbol{\gamma}}$ into Eq. (6) and the fact that $||\boldsymbol{y} - \boldsymbol{\Psi}\boldsymbol{s} - \hat{\boldsymbol{\gamma}}||_2^2 = ||\mathsf{Proj}_{\boldsymbol{\Psi}}(\boldsymbol{\Psi}\boldsymbol{s}) - \mathsf{Proj}_{\boldsymbol{\Psi}}(\boldsymbol{y} - \hat{\boldsymbol{\gamma}})||_2^2$, we have that Θ could be solved from Eq. (6). □

TABLE 7: Grid search range for competitors

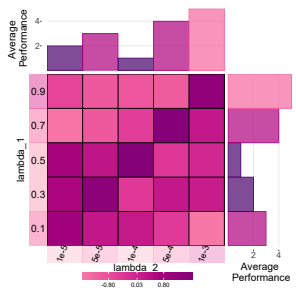
Algorithm	Grid Search Range
Logistic	C={0.05, 0.1, 0.5, 1, 5, 10}
LS-with γ	λ_1, λ_2={0.05, 0.1, 0.5, 1, 5, 10}
Logitstic-with γ	λ_1, λ_2={0.05, 0.1, 0.5, 1, 5, 10}
RankNet	lr={0.001, 0.005, 0.01, 0.05, 0.1} weight_decay={0.0001, 0.001, 0.01, 0.1}
RankSVM	C={0.05, 0.1, 0.5, 1, 5, 10}
RankBoost	n_estimator, n_threshold={10, 20, 30, 40, 50}
GBDT/DART	lr={0.01, 0.05, 0.1, 0.5, 1, 5} n_estimator={20, 30, 40, 50, 60} feature_fraction ={0.7, 0.8, 0.9, 1} bagging_fraction ={0.7, 0.8, 0.9, 1} drop_rate={0.05, 0.1, 0.2, 0.3} skip_drop={0.3, 0.4, 0.5, 0.6, 0.7} bagging_freq={1, 5, 10} max_depth={5, 6, 7, 8, 9}
URLR	mu={0.001, 0.01, 0.05, 0.1, 0.5, 1, 5, 10, 15, 100} p_ratio=[0.05:0.05:0.8]



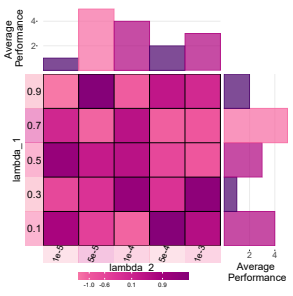
(a) LS-Deep-with γ



(b) Huber-Deep

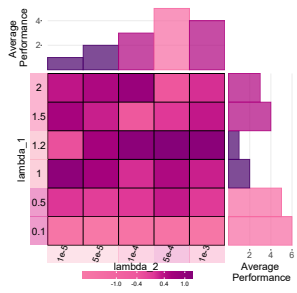


(c) Logit-Deep-with γ

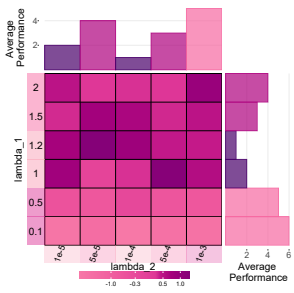


(d) RLogit-Deep

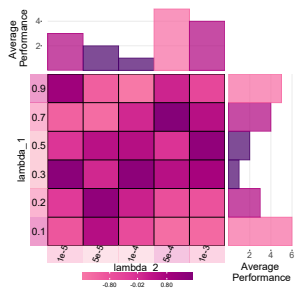
Fig. 10: Average sensitivity analysis on LFW-10 dataset.



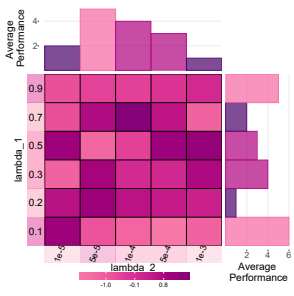
(a) LS-Deep-with γ



(b) Huber-Deep



(c) Logit-Deep-with γ



(d) RLogit-Deep

Fig. 11: Average sensitivity analysis on Shoes dataset.

SENSITIVITY ANALYSIS

LFW-10 Similar to the Human Age Dataset, we show the corresponding fine-grained grid search and the average sensitivity results in Fig.10a-10d, respectively for LS-Deep-with γ, Huber-Deep, Logit-Deep-with γ and RLogit-Deep. We see that these proposed models are not very sensitive towards the parameter perturbation.

Shoes The sensitivity analyses are shown in Fig.11. The results are similar to the first two datasets mentioned above.