Supplementary Materials

1. Ablation Study

In the supplementary materials, 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. We perform the following ablation studies on the three datasets:

- **pretrained+URLR**: In this baseline, we feed the pre-trained feature extracted from Resnet-50 to a traditional robust learning to rank model URLR (a brief introduction of URLR could be found in the main paper). This baseline could show us the power of our method against the pre-trained deep feature.
- **noise+finetuned logit+URLR**: In this baseline, we feed the noisy annotations to 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 noisy data is sufficient for a good feature representation. Moreover, it tells us whether our proposed method outperforms finetuned features learned from noisy labels.
- **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.
- **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 a good feature representation. Moreover, it tells us whether our proposed method outperforms finetuned features learned from voted labels.
- **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.

The ablation results for the three datasets are recorded in Tab.1a-1c, and we have the following findings regarding the results: 1) The finetuned feature merely gains a slight improvement with respect to the pre-trained feature. In fact, without the robust learning mechanism, the vanilla finetuning process (with raw/voting data) could not disentangle the contaminated patterns from the learned features. This weakens the power of traditional robust learning methods (URLR). 2) 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. 3) 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 methods. To sum up, it is vital to do joint end-to-end feature learning and robust ranking.

Table 1: Ablation studies on three datasets.

(a) Ablation studies on Human age dataset.

Algorithm	ACC	F1	Prec.	Rec.	AUC
pretrained+URLR	.7244	.6536	.6381	.6700	.7144
noise+finetuned logit+URLR	.7382	.6733	.6492	.6994	.7319
noise+finetuned l2+URLR	.7380	.6774	.6489	.7086	.7326
major+finetuned logit+URLR	.7391	.6741	.6544	.6949	.7310
major+finetuned l2+URLR	.7381	.6730	.6530	.6943	.7301
LS-Deep-with γ	**.7967**	**.7414**	**.7323**	**.7508**	**.8784**
Logit-Deep-with γ	**.7917**	**.7370**	**.7228**	**.7518**	**.8739**

(b) Ablation studies on Shoes dataset.

Comf.	Fash.	Form.	Pointy	Brown	Open	Ornate	Aver.
.8317	.8299	.8021	.7976	.8042	.7598	.8008	.8037
.8448	.8291	.8030	.8216	.7958	.7278	.8437	.8094
.8492	.8446	.8142	.8011	.8097	.7405	.8358	.8135
.8471	.8434	.8078	.7912	.8268	.7690	.8325	.8168
.8655	.8646	.8294	.8398	.7814	.7217	.7925	.8135
.8500	**.8550**	**.8125**	**.8044**	**.8250**	**.7782**	**.8300**	**.8222**
.8550	**.8500**	**.8200**	**.8339**	**.8125**	**.7481**	**.8325**	**.8217**

(c) Ablation studies on LFW-10 dataset.

Algorithm	Bald	D.Hai	B.Eye	GLook	Masc.	Mouth	Smile	Teeth	Foreh.	Young	Aver.
pre trained+URLR	.5424	.6295	.5213	.6356	.6519	.5699	.6059	.6133	.5746	.6781	.6030
noise +fine tuned logit +URLR	.6695	.6105	.5393	.6059	.6231	.6452	.6373	.6653	.5439	.6802	.6231
noise +fine tuned l2 +URLR	.6568	.6968	.5011	.6377	.6341	.5484	.6059	.6050	.6206	.6781	.6195
major+fine tuned logit+URLR	.6144	.7242	.4989	.6314	.5854	.6301	.6604	.6881	.5987	.6599	.6305
major+fine tuned l2+URLR	.6250	.7495	.5213	.6144	.6009	.6323	.6688	.6445	.6140	.6781	.6361
LS-Deep-with γ	**.6335**	**.7684**	**.5551**	**.6377**	**.6253**	**.7312**	**.7421**	**.7547**	**.6469**	**.7308**	**.6826**
Logit-Deep-with γ	**.6631**	**.7726**	**.5798**	**.6419**	**.5965**	**.7032**	**.7358**	**.7069**	**.6075**	**.6862**	**.6694**