Scale-Varying Triplet Ranking with Classification Loss for Facial Age Estimation

Woobin Im[1] Sungeun Hong[2] Sung-Eui Yoon[1] Hyun S. Yang[1]

[1]Korea Advanced Institute of Science and Technology [2]SK T-Brain

iwbn@kaist.ac.kr csehong@sktbrain.com
sungeui@kaist.edu hsyang@cs.kaist.ac.kr

Abstract. In recent years, considerable efforts based on convolutional neural networks have been devoted to age estimation from face images. Among them, classification-based approaches have shown promising results, but there has been little investigation of age differences and ordinal age information. In this paper, we propose a ranking objective with two novel schemes jointly performed with an age classification objective to take ordinal age labels into account. We first introduce relative triplet sampling in which a set of triplets is constructed considering the relative differences in ages. This also addresses the problem of having limited triplet candidates, that occurs in conventional triplet sampling. We then propose the scale-varying ranking constraint, which decides the importance of a relative triplet and adjusts a scale of gradients accordingly. Our adaptive ranking loss with relative sampling not only lowers the generalization error but ultimately has a meaningful performance improvement over the state-of-the-art methods on two well-known benchmarks.

1 Introduction

There has been a growing interest in age estimation from face images due to a variety of potential applications [1–3]. As in other computer vision fields [4–8], considerable efforts based on Convolutional Neural Networks (CNN) have been devoted to age estimation. Depending on tasks, age estimation can be largely divided into classification of age groups or direct prediction of age values, i.e. the regression task.

In the field of age estimation, CNNs have been widely exploited in a variety of different approaches. To classify age groups, Levi et al. [9] used vanilla CNN with N-class probability outputs, which gives a baseline performance on Adience benchmark dataset [10]. To better estimate ages from face images, studies using transferred CNN [2] and attention models [3] have also been proposed. Meanwhile, studies have been conducted to predict age values beyond the age group classification. Early investigations involved a three-layer CNN regression model with a Gaussian loss [1]. However, recent experiments have shown that training a CNN directly for regression loss is unstable since outliers cause a larger generalization error [11]. This led to different approaches to estimate age values such as distribution-related loss [12–15], ordinal ranking strategy [16, 17],

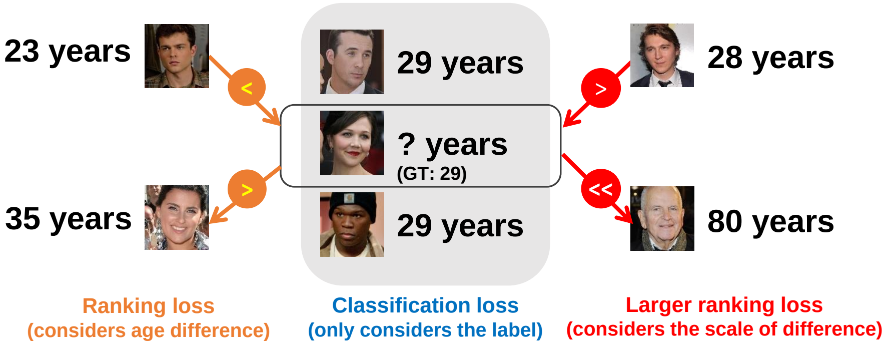


Fig. 1: When we infer the age of the woman in the center, (center) classification loss considers only its label, (left) ranking loss considers the age difference of a triplet, which is an additional clue for inferencing the age, and (right) our adaptive triplet ranking loss considers the scale of differences, so that larger ranking loss is applied to the triplet.

bias-analysis [18], and classification loss [11, 3]. Among them, methods based on classification [11, 3] showed the most simple yet powerful results in large scale datasets in the wild.

Crucially, the classification loss, i.e. cross-entropy loss, however, does not reflect the ordinal characteristics of age labels; it focuses on whether the predicted label is correct, but does not care about the degree of error between a prediction and its target value. As discussed later in the experiment, this leads to a large performance gap between training and validation sets. To address the issue, we take a feature learning approach by an end-to-end learning objective for CNN, which is configured jointly from the proposed ranking constraint as well as the classification loss. The classification loss is used to predict the exact age, while our adaptive ranking constraint, inspired by the triplet ranking loss [19–21] and classification-ranking joint loss [22], acts like a regularizer and consequently helps improve the performance. Meanwhile, large-margin softmax loss [23, 24] is suggested to make the conventional classification loss produce more discriminative feature space which results in better classification performance. However, the approach can be applied when each inter-class relation is the same throughout all pairs of classes; while age pairs have different relations in themselves.

The main difference between conventional triplet loss and our proposed ranking constraint is twofold: relative triplet sampling and scale-varying ranking. Generally, in the conventional triplet loss, triplets consist of two samples with the same label (anchor and positive) and a sample with a different label (negative), and the loss aims to separate the positive pair from the negative pair by a constant margin in the embedding space.

We, however, argue that applying ranking loss by using a constant margin in age estimation cannot fully exploit the ordinal information in age labels.

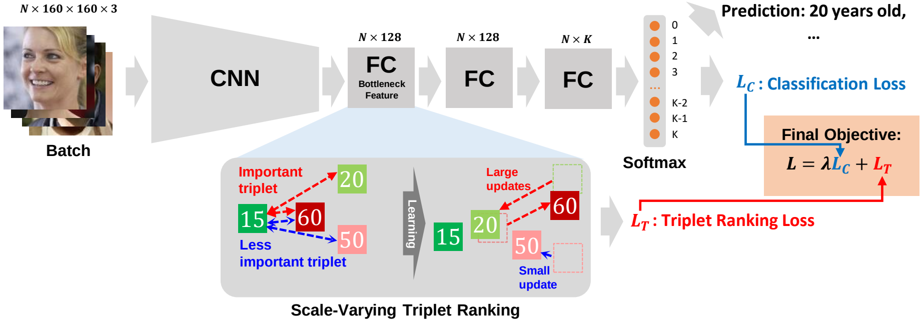


Fig. 2: Overall network framework of our method. In the bottleneck layer, we apply the adaptive triplet ranking strategy (L_T: Eq. 6) by selecting triplets and computing the scale-varying triplet ranking loss. Our final objective jointly includes both the ranking (L_T: Eq. 6) and classification (L_C: Eq. 9) losses simultaneously.

To solve this problem, we first alleviate the existing rigid selection criterion by suggesting relative triplet sampling, where a sample relatively close to the anchor is positive, otherwise negative. The proposed sampling creates more diversity in the triplets than the conventional one, and ultimately makes it possible to apply the following ranking constraint efficiently.

Once the relative triplets are sampled, we then apply the scale-varying ranking loss which automatically decides the importance of a triplet and accordingly adjusts scales of gradients. This enables for a model to learn a ranking without a fixed margin constant and also act like a regularizer, which prevents overfitting of a model. Fig. 1 illustrates the concept and purpose of the proposed method.

The main contributions of this study are as follows: (i) We propose an adaptive, scale-varying ranking loss that prevents overfitting of a model by acting as a regularizer, while assisting in the improvement of the estimation performance. To our knowledge, this is the first attempt to utilize a triplet ranking method to efficiently train a model for age estimation. (ii) To address the lack of possible triplets caused by the conventional triplet sampling, we suggest the relative triplet sampling which also aids the successful application of the scale-varying ranking loss. (iii) We perform extensive experiments in two well-known benchmarks and show meaningful improvement over the state-of-the-art methods, which demonstrates the efficiency of joint training of our ranking loss and the classification objective.

2 Triplet Ranking with Classification

Our method is based on an end-to-end trainable deep convolutional neural network (see Fig. 2), which has the scale-varying triplet ranking module and a softmax output. In the network, our final goal is to estimate a correct age by the softmax layer when a face image is given. While not directly related to the age inference, the triplet ranking module accommodates the relative age difference given a triplet, leading to better age estimation. As a result, our final objective function includes both triplet ranking and classification loss. In the next sub-sections, we introduce our suggested loss functions in detail.

2.1 Relative Triplet Sampling

Sampling triplets is an essential part of a triplet ranking loss. Conventional applications utilizing a triplet loss deal with binary labels, i.e., whether or not two samples belong to the same class. In other words, triplet samples, (a, p, n), commonly called an anchor, a positive, and a negative samples, are chosen, when a and p are in the same class, but a and n are not.

While ages of two faces can be treated as the same or not, we found that it is less effective for ordinal classes like age. One aspect is that the pool of possible triplets in this perspective is restricted. Suppose that we have a mini-batch of size N with an equal number of samples from each class and we have K classes of age labels. If we constrain the positive sample to have the same age label as the anchor for the conventional ranking loss, the pool size of the triplets for a mini-batch becomes $O(N^3/K)$. Since K can be large for an age regression task, e.g., MORPH dataset has 60 classes of age, this approach is subject to severely limited combinations of triplets.

When it comes to age, we can better define the positive and the negative samples by a relative measure. Formally, we sample features from a d-dimensional embedding space in $\mathbb{R}^d$, which is built by a CNN, f, embedding an image input x into $f(x) \in \mathbb{R}^d$. Assume that we have a mini-batch X of a size N with its corresponding set of age labels Y, which contains positive real numbers; i.e., $X = \{x_1, x_2, \cdots, x_N\}$ and $Y = \{y_1, y_2, \cdots, y_N\}$. We then sample every possible $(f(x_a), f(x_p), f(x_n))$, simply dented by (f_a, f_p, f_n), such that the relative triplet satisfies $|y_a - y_p| < |y_a - y_n|$. In other words, the set of our chosen triplets is:

$$\mathcal{T} = \{(f_a, f_p, f_n) \,|\, a \neq p \neq n \,\cap\, |y_a - y_p| < |y_a - y_n|\}. \tag{1}$$

As a result, the chosen relative triplets satisfy that the age difference between the pair of the anchor and the positive must be less than the one between the anchor and the negative. This approach creates more diversity in the triplets compared to the traditional methods, since it has a pool of $O(N^3)$ triplets, which is K times diverse than the conventional one. When used with our adaptive ranking loss, this in turn results in better performance (Table. 1a) and embedding space (Fig. 5).

2.2 Scale-Varying Triplet Ranking Loss

When a triplet ranking is used for representation learning [25, 26, 19], its loss formulation directly utilizes a distance function. For instance, [25] used the squared

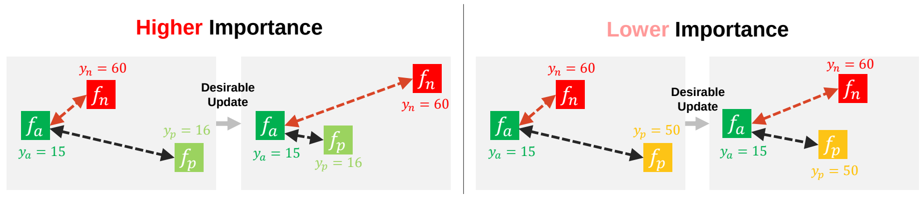


Fig. 3: Schematic visualization of two dimensional embedding space (bottleneck) where similar age samples should be located closer. The left triplet shows a wider discrepancy between age labels and their features in the space, compared to that of the right one. The left triplet should be treated more importantly with its update for learning the feature.

L2 distance between two features:

$$L = \sum_{\mathcal{S}} \max(0, d(f_a, f_p) - d(f_a, f_n) + m), \tag{2}$$

where m is a margin constant, and $d(a, b) = \|a - b\|_2^2$. This loss targets a goal in which the difference between $d(f_a, f_p)$ and $d(f_a, f_n)$ should be larger than m.

Unfortunately, this approach requires the margin constant, and fixing m as a constant for a diverse set of triplets can limit the effectiveness of this strategy. This ineffectiveness is caused, mainly because age triplets may have different importance for learning the feature space – some triplets need a larger m, while others need a smaller m, which is conceptually visualized in Fig. 3.

To design a loss that considers discrepancies in triplets, we propose to use the cross-entropy loss for relative triplets, by normalizing the difference of distances with the softmax function. It enables us to use a loss function, the scale-varying ranking loss, considering our relative triplets, without the margin constant used in the conventional ranking loss.

To compute the loss, we start with a set of relative triplets $\mathcal{T}$. Given $\mathcal{T}$, we calculate a normalized version of positive distance, d_+, and negative distance, d_-. Inspired by [22] we normalize the distances as the following:

$$d_+ = \frac{\exp(d(f_a, f_p))}{\exp(d(f_a, f_p)) + \exp(d(f_a, f_n))}, \quad d_- = \frac{\exp(d(f_a, f_n))}{\exp(d(f_a, f_p)) + \exp(d(f_a, f_n))}. \tag{3}$$

Considering that d_+ and d_- are softmax outputs, we apply the cross-entropy loss for the relative triplet as:

$$l_T(d_+, d_-) = -t_- \log(d_-) - t_+ \log(d_+) = -\log(d_-), \tag{4}$$

where $(t_+, t_-) = (0, 1)$ are our target values; this results in adjusting our feature space such that d_+ approaches to 0 and d_- to 1.

Triplets chosen from training datasets (Eq. 1) could have largely varying importance in learning features. For example, the triplet on the left in Fig. 3 is more important case than the one on the right, since a desirable update for the former case should be stronger than the latter due to its larger discrepancy. If we simply use the cross-entropy loss (Eq. 4), gradients of these two triplets with varying importance are computed to be the same, which fails to achieve the desirable updates.

To reflect the varying importance of relative triplets, we introduce a non-uniform weighting function, $\omega(\cdot)$, that measures the importance of a triplet, as the following:

$$\omega(f_a, f_p, f_n) = \frac{1+\epsilon}{|\bar{y}_a - \bar{y}_p| + \epsilon} - 1, \tag{5}$$

where ϵ is a small constant for preventing zero-division and $\bar{y}_k = (y_k - Y_{\min})/(Y_{\max} - Y_{\min})$ is a normalized label when the range of age labels in a dataset is $\left[Y_{\min}, Y_{\max}\right]$. We then multiply it directly to the loss function, and the final ranking loss L_T is given by:

$$L_T = \frac{1}{|\mathcal{T}|} \sum_{\mathcal{T}} \omega(f_a, f_p, f_n) \cdot l_T(d_+, d_-). \tag{6}$$

Gradient analysis. Before moving on to our final training objective considering the classification loss, we would like to point out that the proposed loss has the same gradient as that of the conventional ranking loss, but it is different in that the magnitude of our gradients are adjusted according to the importance of relative triplets. Note that the conventional ranking loss (Eq. 2) has its derivative with regard to f_a, f_p, and f_n:

$$\frac{\partial L}{\partial f_a} = \sum_{\mathcal{S}} 2(f_n - f_p),\ \frac{\partial L}{\partial f_p} = \sum_{\mathcal{S}} 2(f_p - f_a),\ \frac{\partial L}{\partial f_n} = \sum_{\mathcal{S}} 2(f_a - f_n), \tag{7}$$

where $\mathcal{S} \subset \mathcal{T}$ and $\mathcal{S}$ includes only triplets whose loss is not zeroed out by $\max(0, \cdot)$, and the derivative equals 0 for $\mathcal{T} - \mathcal{S}$. Note that the margin constant does not have any effect on these gradients. On the contrary, our loss function (6) has its derivative:

$$\frac{\partial L_T}{\partial f_a} = \sum_{\mathcal{T}} \alpha(f_n - f_p), \frac{\partial L_T}{\partial f_p} = \sum_{\mathcal{T}} \alpha(f_p - f_a), \frac{\partial L_T}{\partial f_n} = \sum_{\mathcal{T}} \alpha(f_a - f_n). \tag{8}$$

where $\alpha = 2d_+\omega(f_i, f_j, f_k)$. We can see that the directions of the derivatives of two different loss functions are exactly the same, but the scale of ours are regulated by two values: d_+ and ω. d_+ is moving toward zero during training, and if d_+ becomes near zero, our loss also comes closer to zero. The benefit of this is that d_+ softly slows down the learning when the training is adequately done, without using any hyper-parameter such as the margin constant m. Note that we have ω as well as d_+, both of which together let the gradient scale depend on the discrepancies of triplets – those with higher importance will have larger updates, and those with less importance will get smaller updates.

2.3 Final Training Objective

Our final goal is to estimate an age value, and we thus set our model to have a classification endpoint alongside the ranking part. To use age values for training the classification network, we discretize the age values into K classes. We then apply softmax to our classifier. Specifically, our classifier model has one hidden layer with ReLU activation and a softmax layer after the embedding layer. To formulate the classification loss, we have a classifier g, resulting in our whole model to be $g \circ f$, where $\circ$ indicates the function composition. Since g gives probabilities of an input x belonging to each age class, g satisfies $g(f(x)) \in \mathbb{R}^K$, $g(f(x))_j > 0$, and $\sum_{j=1}^{K} g(f(x))_j = 1$, where the subscript j is used to represent the probability belonging to the j-th class.

We also apply the softmax cross-entropy for the classification objective, the same as for the relative triplet ranking loss. Our final classification loss is then defined as:

$$L_C = -\frac{1}{N} \sum_{i=0}^{N} \sum_{j=0}^{K} t_{ij} \log(g(f(x_i))_j), \tag{9}$$

where N is the batch-size, and t_{ij} is an indicator function that has 1 when x_i belongs to the class j, otherwise 0.

Based on our classification and triplet ranking losses, our final training objective function is defined as $L = \lambda L_C + L_T$, where λ is a constant for controlling the balance between L_T and L_C.

3 Experiments

We evaluate our approach with two popular age estimation databases against two different tasks, age regression and age classification: MORPH Album 2 and Adience datasets for each of the tasks respectively.

3.1 Implementation Details

We base our network model on the recent Inception-ResNet-V1 [27] implemented with Tensorflow. We do not start the training from scratch, since our target benchmark databases are relatively small. Instead, we utilize weights pretrained with MS Celeb 1M [28] or ILSVRC2012 datasets.

When we train our model, we use Adam optimizer [29] with a small learning rate 5×10^{-4}, with a exponential decay. In all experiments, we set λ to 0.01, and ϵ to 0.1. For stopping policy, we utilize a portion of a training set as a validation set, and stop training when the validation accuracy converges. We augment the training set with random cropping and color jittering including brightness, saturation and hue. In the test phase, we do not use random cropping, rather we obtain 10 samples by cropping and flipping four corners and the center of an image. Then, we average the scores of the last layer from all 10 samples to compute a final decision.

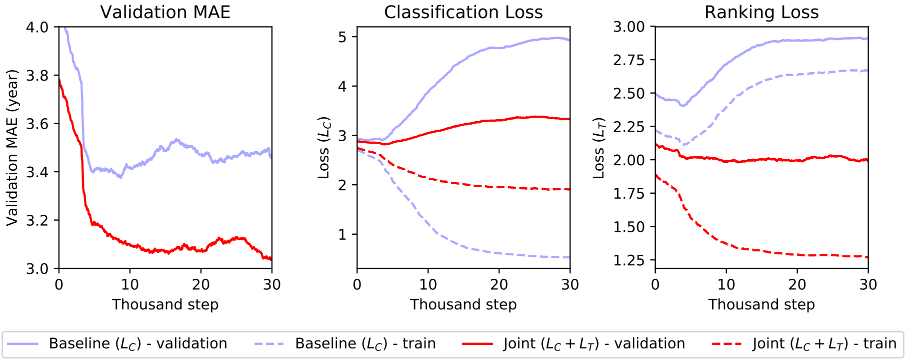


Fig. 4: Training of the baseline and ours on the MORPH Album 2 dataset.

3.2 MORPH Album 2 Dataset

MORPH Album 2 dataset contains 55k facial images of 13k people, and has been widely used by researchers since it provides various labels: identity, gender, age, race, and so on. The MORPH also has been widely used in age estimation field [30, 11, 17]. As Chang et al. suggested, the protocol for evaluation has been settled – using 80% of image samples for training and the rest for testing.

Interestingly, we found that photos of an identity are taken in a short time frame. Specifically, the max age deviation of a identity is only 1.9 years on average. This implies that by perfectly identifying identities, we can achieve down to 1.9 years for the mean absolute error (MAE). In our settings, we also confirmed that using a baseline network pretrained for face verification with MS-Celeb gives MAE of 2.43 years, which is far better than the state-of-the-art result, 2.96 by [17].

Our split for evaluation. To get rid of the effect of identity, we suggest to split the dataset in a way that training and testing sets have no duplicated identities. Thus, we split 13,617 identities into 5 mutually exclusive sets, and perform 5-fold cross-validation for evaluation.

Training-validation curves. Figure 4 shows training-validation curves, with regard to MAE and two types of losses. The first graph plotting MAE, our main target metric, shows a clear gap between the baseline (the solid, light blue curve) and ours (the solid red curve). Especially, we can observe that the baseline overfits in the early stage of training, while our model keeps improving MAE. The second and third plots show that our ranking loss acts as a regularizer that results in lower generalization error to unseen datasets in compensation for a relatively higher training loss than the baseline model.

Comparison between loss types. We report different accuracies obtained by different loss types in Table 1a. The baseline (L_C) does not exploit the ranking loss and has worse MAE than the others. We first compare the baseline to the conventional ranking loss $L_{c.triplet}$ (Eq. 2) adopted from [25] designed for face recognition. Here the results show that the joint loss configuration using classi-

Loss Type	MAE (year)
L_C (Eq. 9)	3.27 ± 0.02
$L_C + L_{\text{c.triplet}}$ (Eq, 2)	2.93 ± 0.01
$L_C + L_T$ (without $\mathcal{T}$)	2.91 ± 0.02
$L_C + L_T$ (**ours**, Eq. 6)	**2.87** $\pm$ 0.02

$\mathcal{T}$: relative triplet selection (Eq. 1)

(a)

Method	MAE (year)
MR-CNN [16]	3.27
DEX [11]	3.25
Ranking-CNN [17]	2.96
Ours (random sample)†	**2.38**
Ours (identity sample)‡	**2.87**

†Pretrained on MS-Celeb, ‡Our split

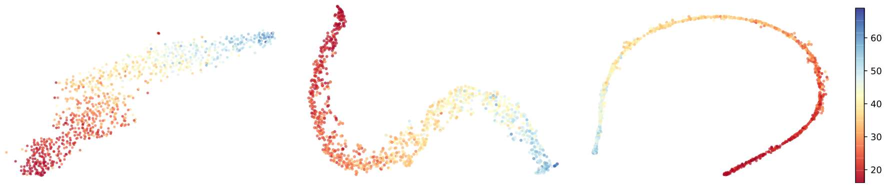
(b)

Table 1: (a) 5-fold cross-validation MAE with the standard error ($\pm e$) on MORPH by our split protocol. We also show the effectiveness of our method over other joint (classification + triplet ranking) losses. (b) Comparison against the state-of-the-art results.

fication loss and ranking loss is effective enough in that they improve MAE in a gap of 0.3 years over the baseline; this supports our fundamental condition that a triplet loss aligns the age feature space better than a classification loss alone does, by utilizing relative information in age labels. As we additionally use the components we designed, performance is further improved. Without exploiting the relative triplet selection, our ranking loss L_T shows performance 0.02 years better than $L_{\text{c.triplet}}$. Furthermore, ours works even better than the other joint models when combined with the relative sampling method, by showing MAE of 2.87, which is the lowest result among all the tested methods. This improvement is mainly resulted from the relative sampling for diverse set of triplets, and our adaptive scale-varying loss function (Eq. 6) leading to reasonable gradients (Eq. 8) for ordinal classes.

Comparisons against the state-of-the-art. In Table 1b, we compare our model to other CNN models. First, we can conclude that if we use a facial domain knowledge, i.e. pretraining on MS-Celeb, we can achieve the highest result based on the previously widely used split protocol, i.e. random split by images [30, 11, 17]. When we use our harder split, i.e. random split by identity, we achieve MAE of 2.87, which is also better than results from the prior state-of-the-art methods.

Embedding space visualization. Fig. 5 visualizes the embedding space computed by only classification loss, joint loss with $L_{c.triplet}$ [25], and our joint model. Here, we can clearly observe that ours (Fig. 5c) much coherently aligns the features along the one dimensional curve as a function of age than the others (Fig. 5a-5b). That is because the classification loss is only aware of the class difference rather than considering the ordinal characteristics; samples in similar colors (and thus ages) as well as those in totally different colors are thus treated equally, resulting in a rather fuzzy feature space. In the joint loss case (Fig. 5b), the samples are aligned in more neat shape, but not in complete 1D curve, since it has a fixed margin term without considering different importance of triplets. On the other hand, the scale-varying ranking loss deliverately put those in sim-



(a) Classification: L_C (b) Joint: $L_C + L_{\text{c.triplet}}$ (c) Ours: $L_C + L_T$

Fig. 5: Embedding space visualization of a bottleneck feature of the network by T-SNE [31] method. Input from test instances of the MORPH database. Values on the color bar are ages.

Method	Exact (%)	1-off (%)
CNN [9]	50.7 ± 5.1	84.7 ± 2.2
DEX [11]	55.7 ± 6.1	89.7 ± 1.8
Attention CNN [3]	61.8 ± 2.1	95.1 ± 0.03
Squared EMD [13]	62.2	94.3
Baseline (L_C)	60.5 ± 2.2	95.0 ± 0.6
Ours ($L_C + L_T$)	$\mathbf{63.1 \pm 1.0}$	$\mathbf{96.7 \pm 0.4}$

Table 2: Comparison to the state-of-the-art deep methods on the Adience benchmark. '1-off' means that 1-off class miss classification is allowed as correct. For 'exact' results, we do not allow any mis-classification. Alongside the accuracy, we report the standard error ($\pm e$) of 5-fold cross-validation results.

ilar colors at close locations while those in different colors are pushed farther, considering how close or far they should be.

3.3 Adience Benchmark

We also evaluate our model to the age classification task using the Adience benchmark database [10]. The database includes 25k cropped face images taken in unconstrained environments. It provides identity, gender, and age group labels for each face image. For performance evaluation, we follow the protocol used by [9]. The dataset consists of 5 splits where 5-fold cross-validation is performed. Its age groups include eight classes: $[0, 2]$, $[4, 6]$, $[8, 12]$, $[15, 20]$, $[25, 32]$, $[38, 43]$, $[48, 53]$, and $[60, 100]$.

Performance analysis. We report age classification results and compare our results to other methods in Table 2. For a baseline, we first train our baseline model with only classification loss, which produces 60.5% of accuracy. When we train the network with our method, we can clearly see the improvement of ours over the baseline with about 3% of gap in exact and 2% in 1-off results. In regard to the fact that other work [9, 11, 3] use L_C (Eq. 9) for classification, we can expect that adding our adaptive ranking loss (L_T: Eq. 6) to their classification loss (L_C: Eq. 9) is able to further improve the performance. This concludes that our scale-varying triplet loss acts as a reasonable objective function whether labels are highly dense (i.e. regression), or not (i.e. classification).

4 Conclusion

We have proposed the adaptive, scale-varying ranking loss jointly used with the classification loss for age estimation. Based on a simple intuition that a triplet ranking loss is helpful for age feature learning, we adapt the conventional one by introducing the relative triplet selection and the weighting scheme to improve the performance of the joint objective for age estimation. By using our proposed joint loss with the relative triplet sampling, we show that our adaptive scale-varying ranking loss reduces the generalization error of a model and better aligns the age features than the baseline. Lastly, our approach achieves meaningful improvements over the state-of-the-art methods in both age regression and classification tasks.

Much interesting future work lies ahead. While the proposed approach was applied mainly to the estimation of facial age in this study, it is not restricted only to this tested application. Since our work uses a relative ranking strategy, it can be applied to other domains where a distance measure between ground-truth labels exists.

Acknowledgement. This work was supported by Institute for Information & Communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) [2016-0-00562(R0124-16-0002), Emotional Intelligence Technology to Infer Human Emotion and Carry on Dialogue Accordingly], and NRF [NRF-2017M3C4A7066317].

References

1. Ranjan, R., Zhou, S., Cheng Chen, J., Kumar, A., Alavi, A., Patel, V.M., Chellappa, R.: Unconstrained age estimation with deep convolutional neural networks. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. (2015) 109–117
2. Ozbulak, G., Aytar, Y., Ekenel, H.K.: How transferable are cnn-based features for age and gender classification? In: Biometrics Special Interest Group (BIOSIG), 2016 International Conference of the, IEEE (2016) 1–6
3. Rodríguez, P., Cucurull, G., Gonfaus, J.M., Roca, F.X., Gonzàlez, J.: Age and gender recognition in the wild with deep attention. Pattern Recognition (2017)
4. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. (2012) 1097–1105
5. Hong, S., Ryu, J., Im, W., Yang, H.S.: D3: Recognizing dynamic scenes with deep dual descriptor based on key frames and key segments. Neurocomputing **273** (2018) 611–621
6. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2014) 1701–1708

7. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Web-scale training for face identification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2746–2754
8. Hong, S., Im, W., Ryu, J., Yang, H.S.: Sspp-dan: Deep domain adaptation network for face recognition with single sample per person. In: International Conference on Image Processing. (2017)
9. Levi, G., Hassner, T.: Age and gender classification using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2015) 34–42
10. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security **9** (2014) 2170–2179
11. Rothe, R., Timofte, R., Van Gool, L.: Deep expectation of real and apparent age from a single image without facial landmarks. International Journal of Computer Vision (2016) 1–14
12. Geng, X., Yin, C., Zhou, Z.H.: Facial age estimation by learning from label distributions. IEEE transactions on pattern analysis and machine intelligence **35** (2013) 2401–2412
13. Hou, L., Yu, C.P., Samaras, D.: Squared earth mover's distance-based loss for training deep neural networks. arXiv preprint arXiv:1611.05916 (2016)
14. Gao, B.B., Xing, C., Xie, C.W., Wu, J., Geng, X.: Deep label distribution learning with label ambiguity. IEEE Transactions on Image Processing **26** (2017) 2825–2838
15. Gao, B.B., Zhou, H.Y., Wu, J., Geng, X.: Age estimation using expectation of label distribution learning. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization (2018) 712–718
16. Niu, Z., Zhou, M., Wang, L., Gao, X., Hua, G.: Ordinal regression with multiple output cnn for age estimation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. (2016) 4920–4928
17. Chen, S., Zhang, C., Dong, M., Le, J., Rao, M.: Using ranking-cnn for age estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
18. Clapés, A., Bilici, O., Temirova, D., Avots, E., Anbarjafari, G., Escalera, S.: From apparent to real age: gender, age, ethnic, makeup, and expression bias analysis in real age estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2018) 2373–2382
19. Wang, J., Song, Y., Leung, T., Rosenberg, C., Wang, J., Philbin, J., Chen, B., Wu, Y.: Learning fine-grained image similarity with deep ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2014) 1386–1393
20. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: BMVC. Volume 1. (2015) 6
21. Hong, S., Im, W., Yang, H.S.: Cbvmr: Content-based videomusic retrieval using soft intra-modal structure constraint. In: ACM International Conference on Multimedia Retrieval. (2018)
22. Simo-Serra, E., Ishikawa, H.: Fashion style in 128 floats: Joint ranking and classification using weak data for feature extraction. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 298–307
23. Liu, W., Wen, Y., Yu, Z., Yang, M.: Large-margin softmax loss for convolutional neural networks. In: ICML. (2016) 507–516

24. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Volume 1. (2017) 1
25. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 815–823
26. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 5005–5013
27. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI. (2017) 4278–4284
28. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A dataset and benchmark for large scale face recognition. In: European Conference on Computer Vision, Springer (2016)
29. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014)
30. Chang, K.Y., Chen, C.S., Hung, Y.P.: Ordinal hyperplanes ranker with cost sensitivities for age estimation. In: Computer vision and pattern recognition (cvpr), 2011 ieee conference on, IEEE (2011) 585–592
31. Maaten, L.v.d., Hinton, G.: Visualizing data using t-sne. Journal of Machine Learning Research **9** (2008) 2579–2605