# Supplementary for UniTSFace

**Qiufu Li**
liqiufu@szu.edu.cn

**Xi Jia**
x.jia.1@cs.bham.ac.uk

**Jiancan Zhou**
zhoujiancan@foxmail.com

**Linlin Shen**
llshen@szu.edu.cn

**Jinming Duan**
j.duan@bham.ac.uk

## 1 Inequalities about $L_{\text{naive}}$

We here present the detailed derivations for the inequalities about the naive loss $L_{\text{naive}}$. For any sample $\boldsymbol{X}^{(i)}$ captured from subject $i$, with feature $\boldsymbol{x}^{(i)} = \mathcal{M}(\boldsymbol{X}^{(i)})$, the naive loss $L_{\text{naive}}$ comprises of one positive sample-to-sample similarity and $N-1$ negative similarities,

$$L_{\text{naive}}(\boldsymbol{X}^{(i)}) = -\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) + \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)}), \tag{1}$$

where $\boldsymbol{x}_*^{(i)} = \mathcal{M}(\boldsymbol{X}_*^{(i)}),\ \boldsymbol{x}_*^{(j)} = \mathcal{M}(\boldsymbol{X}_*^{(j)})$, and $\boldsymbol{X}_*^{(i)}$ and $\boldsymbol{X}_*^{(j)}$ are randomly taken from the subject $i, j$, with $j \neq i$.

Using the inequality of arithmetic and geometric means[1], we derive four inequalities about $L_{\text{naive}}$.

$$\begin{aligned}
& L_{\text{naive}}(\boldsymbol{X}^{(i)}) \\
= \ & -\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) + \frac{1}{N-1}\sum_{\substack{j=1\\ j\neq i}}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)}) \\
= \ & -\frac{N}{N-1}\Big(\frac{N-1}{N}\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) - \frac{1}{N}\sum_{\substack{j=1\\ j\neq i}}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\Big) && (2) \\
= \ & -\frac{N}{N-1}\Big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) - \frac{1}{N}\sum_{j=1}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\Big) && (3) \\
= \ & -\frac{N}{N-1}\Big[\log\exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big) - \log\exp\Big(\frac{1}{N}\sum_{j=1}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\Big)\Big] && (4) \\
\leq \ & -\frac{N}{N-1}\Big[\log\exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big) - \log\Big(\frac{1}{N}\sum_{j=1}^{N} \exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\big)\Big)\Big] && (5) \\
= \ & -\frac{N}{N-1}\Big(\log\frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}}{\sum_{j=1}^{N}\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})}} + \log N\Big) && (6) \\
= \ & -\frac{N}{N-1}\log\frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}}{\sum_{j=1}^{N}\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})}} - \frac{N\log N}{N-1}, && (7)
\end{aligned}$$

[1] $\sqrt[n]{\prod_{i=1}^{n} a_i} \leq \frac{1}{n}\sum_{i=1}^{n} a_i$ for $a_i \geq 0$.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

$$
\begin{aligned}
& L_{\text{naive}}(\boldsymbol{X}^{(i)}) \\
= \; & -\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) + \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)}) \\
= \; & \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) + \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\big) - 2\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) && (8) \\
= \; & 2\Big[\frac{1}{2}\Big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) + \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\Big) - \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\Big] && (9) \\
= \; & 2\Big\{ \log\exp\Big[\frac{1}{2}\Big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) + \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\Big)\Big] - \log\exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big)\Big\} && (10) \\
\leq \; & 2\Big\{ \log\Big[\frac{1}{2}\Big( \exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big) + \exp \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\big)}{N-1}\Big)\Big] - \log\exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big)\Big\} && (11) \\
= \; & 2\Big[ \log\Big( \exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big) + \exp \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})\big)}{N-1}\Big) - \log\exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big) - \log 2\Big] && (12) \\
= \; & 2\log\Big(1 + \frac{\exp\big(\sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})}{N-1}\big)}{\exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big)}\Big) - 2\log 2, && (13)
\end{aligned}
$$

$$
\begin{aligned}
& L_{\text{naive}}(\boldsymbol{X}^{(i)}) \\
= \; & -\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)}) + \frac{1}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)}) \\
= \; & \frac{1}{N-1}\Big( \sum_{\substack{j=1 \\ j \neq i}}^{N} \big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)}) + \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big) - 2(N-1)\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\Big) && (14) \\
= \; & \frac{2}{N-1}\Big( \sum_{\substack{j=1 \\ j \neq i}}^{N} \frac{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)}) + \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}{2} - (N-1)\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\Big) && (15) \\
= \; & \frac{2}{N-1}\Big( \sum_{\substack{j=1 \\ j \neq i}}^{N} \log\exp \frac{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)}) + \gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}{2} - (N-1)\log\exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big)\Big) && (16) \\
\leq \; & \frac{2}{N-1}\Big( \sum_{\substack{j=1 \\ j \neq i}}^{N} \log \frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})} + \mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}}{2} - (N-1)\log\exp\big(\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})\big)\Big) && (17) \\
= \; & \frac{2}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \Big( \log \frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})} + \mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}}{2} - \log \mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}\Big) && (18) \\
= \; & \frac{2}{N-1} \sum_{\substack{j=1 \\ j \neq i}}^{N} \log \frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(j)})} + \mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}}{2\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)}, \boldsymbol{x}_*^{(i)})}} && (19)
\end{aligned}
$$

$$= \frac{2}{N-1}\sum_{\substack{j=1\\j\neq i}}^{N}\log\Big[\frac{1}{2}\Big(1+\frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})}}{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(i)})}}\Big)\Big] \tag{20}$$

$$= \frac{2}{N-1}\sum_{\substack{j=1\\j\neq i}}^{N}\log\Big(1+\frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})}}{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(i)})}}\Big)-2\log 2, \tag{21}$$

and

$$\sum_{i=1}^{N} L_{\text{naive}}(\boldsymbol{X}^{(i)})$$

$$= \frac{1}{N-1}\Big(\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})-(N-1)\sum_{i=1}^{N}\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(i)})\Big) \tag{22}$$

$$= \frac{1}{N-1}\Big(\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})-\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})\Big) \tag{23}$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\Big(\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})-\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})\Big) \tag{24}$$

$$= \frac{1}{N-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\Big(\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})+\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})-2\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})\Big) \tag{25}$$

$$= \frac{2}{N-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\Big(\log\exp\frac{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})+\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})}{2}-\log\exp\big(\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})\big)\Big) \tag{26}$$

$$\leq \frac{2}{N-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\Big(\log\frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})}+\mathrm{e}^{\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})}}{2}-\log\exp\big(\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})\big)\Big) \tag{27}$$

$$= \frac{2}{N-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\Big[\log\Big(\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})}+\mathrm{e}^{\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})}\Big)-\log\exp\big(\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})\big)\Big]-2N\log 2 \tag{28}$$

$$= \frac{2}{N-1}\sum_{i=1}^{N}\sum_{\substack{j=1\\j\neq i}}^{N}\log\Big(1+\frac{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})}}{\mathrm{e}^{\gamma g(\boldsymbol{x}^{(j)},\boldsymbol{x}_*^{(j)})}}\Big)-2N\log 2. \tag{29}$$

## 2 Marginal Sample-to-Sample Based Losses

We have derived three sample-to-sample based losses in the manuscript, i.e., USS loss, sample-to-sample based softmax, and BCE losses. We hereby present their respective marginal versions:

$$L_{\text{uss-m}}(\boldsymbol{X}^{(i)}) = \log\Big(1+\mathrm{e}^{-\gamma(g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(i)})-m)+b}\Big)+\sum_{\substack{j=1\\j\neq i}}^{N}\log\Big(1+\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})-b}\Big), \tag{30}$$

$$L_{\text{soft-m}}(\boldsymbol{X}^{(i)}) = = -\log\frac{\mathrm{e}^{\gamma(g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(i)})-m)}}{\sum_{j=1}^{N}\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})}}, \tag{31}$$

$$L_{\text{bce-m}}(\boldsymbol{X}^{(i)}) = \log\Big(1+\mathrm{e}^{-\gamma(g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(i)})-m)+b_i}\Big)+\sum_{\substack{j=1\\j\neq i}}^{N}\log\Big(1+\mathrm{e}^{\gamma g(\boldsymbol{x}^{(i)},\boldsymbol{x}_*^{(j)})-b_j}\Big). \tag{32}$$

The experimental evaluations of such marginal losses have been included in Sec. 4.3 of the manuscript.

## 3 Training Details

In our work, we choose the cosine function to represent the similarity of two features, i.e.,

$$g(\boldsymbol{x}, \boldsymbol{x}_*) = \cos(\boldsymbol{x}, \boldsymbol{x}_*) = \frac{\langle \boldsymbol{x}, \boldsymbol{x}_* \rangle}{\|\boldsymbol{x}\| \|\boldsymbol{x}_*\|}, \quad \text{for } \forall\, \boldsymbol{x}, \boldsymbol{x}_* \in \mathcal{F}. \tag{33}$$

Following (2; 1), we use customized ResNets (such as ResNet-50, ResNet-100, and ResNet-200) as our backbone networks. We implement all models using Pytorch and train them using the SGD optimizer with a weight decay of 5e-4 and momentum of 0.9. We use $\gamma = 64$ for $L_{\text{uss}}$, $L_{\text{soft}}$, and $L_{\text{bce}}$ in all experiments. Note that, we use a combination of CosFace($m = 0.4$) and our USS($m = 0.1$) as UniTSFace in Tables 4 and 5.

For the face models (using ResNet-50 as the backbone) on CASIA-WebFace, we train them over 28 epochs with a batch size of 512. The learning rate starts at 0.1 and is reduced by a factor of 10 at the $16^{th}$ and $24^{th}$ epoch. All models in ablation and parameter study were trained on CASIA-WebFace.

In Table 1, the margin $m$ of $L_{\text{soft-m}}$, $L_{\text{bce-m}}$ and $L_{\text{uss-m}}$ are set to be 0.1. In Table 3, the margin $m$ of ArcFace and CosFace are set to be 0.5 and 0.4 respectively. The UniTSFace under the 'Small' protocol of MegaFace Challenge 1 in Table 4 and the models re-implemented in MFR Ongoing (the first ten rows in Table 5) were also trained on CASIA-WebFace.

For Glint360K, we train the models(ResNet-100) for 20 epochs using a batch size of 1024. Initially, the learning rate was set at 0.1, and a polynomial decay strategy (power=2) was applied to the learning rate schedule. The UniTSFace under the 'Large' protocol of MegaFace Challenge 1 (as shown in Table 4) was trained on Glint360K.

For WebFace4M, we train the models(ResNet-50) for 20 epochs using a batch size of 1024. The learning rate was initially set at 0.1, while a polynomial decay strategy (power=2) was applied to the learning rate schedule. The UniTSFace at the $12^{th}$ row of Table 5 was trained on WebFace4M.

In the case of WebFace42M, we train the models(ResNet-200) for 20 epochs, using a larger batch size of 4096. The learning rate linearly warmed up from 0 to 0.4 during the first epoch, followed by a polynomial decay (power=2) for the remaining 19 epochs. The UniTSFace at the last row of Table 5 was trained on WebFace42M.

## 4 Megaface Challenge 1

We notice that the official MegaFace challenge website has been decommissioned and MegaFace data are no longer being distributed. However, despite these changes, we can still utilize the previously released data and development kit to evaluate the performance of trained models. Specifically, we have adopted the MegaFace testsuite provided by InsightFace (`https://github.com/deepinsight/insightface/tree/master/recognition/_evaluation_/megaface`), which also includes the official devkit.

## References


[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019.

[2] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.