Rethinking Feature-based Knowledge Distillation for Face Recognition
– Supplementary Material

Jingzhi Li†,*,1, Zidong Guo*,1, Hui Li1, Seungju Han2, Ji-won Baek2, Min Yang1, Ran Yang1, Sungjoo Suh2
1Samsung R&D Institute China Xi’an (SRCX)
2Samsung Advanced Institute of Technology (SAIT), South Korea
jingzhi.li, zidong.guo, hui01.li, sj75.han, jw0328.baek, minl6.yang,
ran01.yang, sungjoo.suh@samsung.com

1. Introduction
In this supplementary material, we first present the details of the estimation of the intrinsic dimension. Then to affirm the universality of tailored teachers, we include additional experiments on ReFO+. Finally, we present the effect of using different β in reverse distillation and feature-only distillation to show the insensitivity on changes in distillation weight.

2. Estimating Intrinsic Dimension with TwoNN
To estimate the intrinsic dimension of the network, we follow the TwoNN method [2] as applied in [1].

Theory. TwoNN is a global intrinsic dimension estimator based on the distances of the first two nearest neighbors of each point in the space. Let \( r_1 \) and \( r_2 \) denote the distances to the nearest and the second nearest neighbors. The volume of the hyperspherical shell enclosed by the two neighbors is related to the intrinsic dimension by,

\[
\Delta v = w_d (r_2^d - r_1^d),
\]

where \( d \) is the intrinsic dimension and \( w_d \) is the volume of a unit \( d \)-sphere. It is proven in [2] that for uniformly sampled points, the ratio \( \mu = \frac{r_2}{r_1} \) follows Pareto distribution with parameter \( d + 1 \) as,

\[
f(\mu|d) = d \mu^{-(d+1)},
\]

\( d \) can then be simply computed by maximizing the likelihood,

\[
P(\mu|d) = d^N \prod_{i=1}^{N} \mu_i^{-(d+1)},
\]

where \( \mu = (\mu_1, \mu_2, \ldots, \mu_N) \) is the vector of sampled ratios.

3. Universality of ReFO+
In the main text, we have shown the universality of ReFO using two IR100 teachers tailored to IR18 and MFN respectively. In Tab. 1, we present the same set of experiments using ReFO+. It is clear that both tailored teachers bring performance improvements on students of different structure. Comparing to ReFO, all students from ReFO+ show higher

<table>
<thead>
<tr>
<th>Teacher</th>
<th>Student</th>
<th>MFN</th>
<th>MNv2</th>
<th>IR18</th>
<th>IR34</th>
</tr>
</thead>
<tbody>
<tr>
<td>IR100</td>
<td></td>
<td>53.86</td>
<td>58.32</td>
<td>61.70</td>
<td>73.16</td>
</tr>
<tr>
<td>IR100( \text{ReFO} + )</td>
<td>MFN</td>
<td>58.66</td>
<td>64.20</td>
<td>67.26</td>
<td>76.40</td>
</tr>
<tr>
<td>IR100( \text{ReFO} + )</td>
<td>IR18</td>
<td>59.64</td>
<td>64.53</td>
<td>68.56</td>
<td>77.38</td>
</tr>
</tbody>
</table>

*Equal contribution. † Corresponding author.
accuracies with a mean absolute performance improvement of 1.4% on top of students trained by ReFO.

4. Effects of $\beta_1$ in Reverse Distillation

In reverse distillation (Sec.3.3, Algorithm 1, Step 3), the teacher is trained by the loss

$$L = L_{cls} + \beta_1 L_{emb}(f_t, f_s).$$

(4)

There is a balancing weight $\beta_1$ controlling the emphasis on student feature space awareness. We generally keep it small to allow the teacher to focus on the optimization of the main task. The default value of $\beta_1$ is 0.5 for normalized embeddings and 0.001 for un-normalized embeddings. In Tab. 2, we show the results of using other values of $\beta_1$ on the IR100-IR18 pair with normalized embeddings, evaluated on the three tracks of ICCV21-MFR. On the largest MR-all track, all values of $\beta_1$ are rather comparable with $\beta_1 = 1.0$ perform slightly better. The smaller Mask track and Children track show more variations in performance, and we opt for $\beta_1 = 0.5$ for its better overall results.

Table 2. Accuracies of different $\beta_1$ during reverse distillation for IR18 on ICCV21-MFR (%). Teacher: IR100. The best results are in bold.

<table>
<thead>
<tr>
<th>Track</th>
<th>$\beta_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>MR-all</td>
<td>66.09</td>
</tr>
<tr>
<td>Mask</td>
<td>46.10</td>
</tr>
<tr>
<td>Children</td>
<td>39.42</td>
</tr>
</tbody>
</table>

5. Effects of $\beta_2$ in Feature-Only Distillation

In the second stage of feature-only distillation (Sec.3.3, Algorithm 1, Step 5), the final student is trained by the loss

$$L = \beta_2 L_{emb}(f_s, f_t).$$

(5)

There is only one parameter $\beta_2$. In principle changing its value is equivalent to simultaneously adjusting the learning rate and the weight decay parameter. To keep the optimization parameters consistent across experiments, we allow $\beta_2$ to change instead. In Tab. 3, we show the results of ReFO+ with different choices of $\beta_2$ on two teacher-student pairs, IR50-MFN and IR100-IR18. While both teacher-student pairs show the best performance at $\beta_2 = 5$, the IR100-IR18 pair appears more insensitive to different values of $\beta_2$.

Table 3. Accuracies of different $\beta_2$ during FO distillation on MR-all (%). The best results are in bold.

<table>
<thead>
<tr>
<th>Model</th>
<th>$\beta_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
</tr>
<tr>
<td>MFN</td>
<td>58.11</td>
</tr>
<tr>
<td>IR18</td>
<td>68.37</td>
</tr>
</tbody>
</table>

References
