# Teaching Where to Look: Attention Similarity Knowledge Distillation for Low Resolution Face Recognition

Sungho Shin[1[0000−0001−5393−6169]], Joosoon Lee[1[0000−0001−6262−5303]], Junseok Lee[1[0000−0001−5212−2657]], Yeonguk Yu[1[0000−0003−2147−4718]], and Kyoobin Lee[1[0000−0003−4299−4923]] ⋆

School of Integrated Technology (SIT), Gwangju Institute of Science and Technology (GIST), Cheomdan-gwagiro 123, Buk-gu, Gwangju 61005, Republic of Korea.
{hogili89,joosoon1111,junseoklee,yeon_guk,kyoobinlee}
@gist.ac.kr

**Abstract.** Deep learning has achieved outstanding performance for face recognition benchmarks, but performance reduces significantly for low resolution (LR) images. We propose an attention similarity knowledge distillation approach, which transfers attention maps obtained from a high resolution (HR) network as a teacher into an LR network as a student to boost LR recognition performance. Inspired by humans being able to approximate an object's region from an LR image based on prior knowledge obtained from HR images, we designed the knowledge distillation loss using the cosine similarity to make the student network's attention resemble the teacher network's attention. Experiments on various LR face related benchmarks confirmed the proposed method generally improved recognition performances on LR settings, outperforming state-of-the-art results by simply transferring well-constructed attention maps. The code and pretrained models are publicly available in the https://github.com/gist-ailab/teaching-where-to-look.

**Keywords:** Attention similarity knowledge distillation, cosine similarity, low resolution face recognition

## 1 Introduction

Recent face recognition model recognizes the identity of a given face image from the 1M distractors with an accuracy of 99.087% [15]. However, most face recognition benchmarks such as MegaFace [15], CASIA [33], and MS-Celeb-1M [10] contain high resolution (HR) images that differ significantly from real-world environments, typically captured by surveillance cameras. When deep learning approaches are directly applied to low resolution (LR) images after being trained on HR images, significant performance degradation occurred [1, 21, 30].

To overcome the LR problem associated with face recognition, prior knowledge extracted from HR face images is used to compensate spatial information

⋆ Corresponding author.

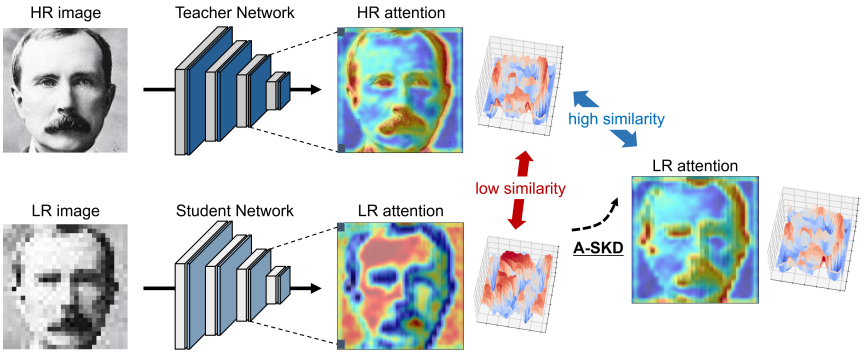


**Fig. 1.** Proposed attention similarity knowledge distillation (A-SKD) concept for low resolution (LR) face recognition problem. Well-constructed attention maps from the HR network are transferred to the LR network by forming high similarity between them for guiding the LR network to focus on detailed parts captured by the HR network. Face images and attention maps are from the AgeDB-30 [23].

loss. Depending on the approach of transferring the prior knowledge to LR image domain, LR face recognition methods are categorized into two types: super-resolution and knowledge distillation based approaches. Super-resolution based approaches utilize generative models to improve LR images to HR before input to recognition networks [7, 9, 11, 16, 28, 30]. Following the development of super-resolution methods, LR images can be successfully reconstructed into HR images and recognized by a network trained on HR images [5, 6, 19, 26]. However, super-resolution models incur high computational costs for both training and inference, even larger than the costs required for recognition networks. Furthermore, generating HR from LR images is an ill-posed problem, i.e., many HR images can match with a single LR image [4]; hence the identity of a LR image can be altered.

To combat this, knowledge distillation based methods have been proposed to transfer prior knowledge from HR images to models trained on LR face images [8, 22, 37]. When the resolution of face images is degraded, face recognition models cannot capture accurate features for identification due to spatial information loss. In particular, features from detailed facial parts are difficult to be captured from a few pixels on LR images, e.g. eyes, nose, and mouth [18]. Previous studies mainly focused on feature based knowledge distillation (F-KD) methods to encourage the LR network's features to mimic the HR network's features by reducing the Euclidean distance between them [8, 22, 37]. The original concept of F-KD was proposed as a lightweight student model to mimic features from over-parameterized teacher models [34]. Because teacher model's features would generally include more information than the student model, F-KD approaches improve the accuracy of the student model. Similarly, informative features from

the HR network are distilled to the LR network in the LR face recognition problems.

This study proposes the attention similarity knowledge distillation approach to distill well-constructed attention maps from an HR network into an LR network by increasing similarity between them. The approach was motivated by the observation that humans can approximate an object's regions from LR images based on prior knowledge learned from previously viewed HR images. Kumar et al. proposed that guiding the LR face recognition network to generate facial keypoints (e.g., eyes, ears, nose, and lips) improved recognition performance by directing the network's attention to the informative regions [18]. Thus, we designed the prior knowledge as an attention map and transferred the knowledge by increasing similarity between the HR and LR networks' attention maps.

Experiments on LR face recognition, face detection, and general object classification demonstrated that the attention mechanism was the best prior knowledge obtainable from the HR networks and similarity was the best method for transferring knowledge to the LR networks. Ablation studies and attention analyses demonstrated the proposed A-SKD effectiveness.

## 2 Related Works

**Knowledge distillation.** Hinton et al. first proposed the knowledge distillation approach to transfer knowledge from a teacher network into a smaller student network [12]. Soft logits from a teacher network were distilled into a student network by reducing the Kullback-Leibler (KL) divergence score, which quantifies the difference between the teacher and student logits distributions. Various F-KD methods were subsequently proposed to distill intermediate representations [25, 27, 34, 35]. FitNet reduced the Euclidean distance between teacher and student network's features to boost student network training [27]. Zagoruyko et al. proposed attention transfer (AT) to reduce the distance between teacher and student network's attention maps rather than distilling entire features [35]. Since attention maps are calculated by applying channel-wise pooling to feature vectors, activation levels for each feature can be distilled efficiently. Relational knowledge distillation (RKD) recently confirmed significant performance gain by distilling structural relationships for features across teacher and student networks [25].

**Feature guided LR face recognition.** Various approaches that distill well-constructed features from the HR face recognition network to the LR network have been proposed to improve LR face recognition performances [8, 22, 37]. Conventional knowledge distillation methods assume that over-parameterized teacher networks extract richer information and it can be transferred to smaller student networks. Similarly, LR face recognition studies focused on transferring knowledge from networks trained on highly informative inputs to networks trained on less informative inputs. Zhu et al. introduced knowledge distillation approach for LR object classification [37], confirming that simple logit distillation from the HR to LR network significantly improved LR classification performance,

even superior to super-resolution based methods. F-KD [22] and hybrid order relational knowledge distillation (HORKD) [8], which is the variant of RKD [25], methods were subsequently applied to LR face recognition problems to transfer intermediate representations from the HR network.

Another approach is to guide the LR network by training it to generate keypoints (e.g. eyes, ears, nose, and lips) [18]. An auxiliary layer is added to generate keypoints, and hence guide the network to focus on specific facial characteristics. It is well known that facial parts such as eyes and ears are important for recognition [17, 18], hence LR face recognition networks guided by keypoints achieve better performance. Inspired by this, we designed the attention distillation method that guides the LR network to focus on important regions of the HR network. However, attention distillation methods have not been previously explored for LR face recognition. We investigated the efficient attention distillation methods for LR settings and proposed the cosine similarity as the distance measure between HR and LR network's attention maps.

## 3 Method

### 3.1 Low resolution image generation

We require HR and LR face image pairs to distill the HR network's knowledge to the LR network. Following the protocol for LR image generation in super-resolution studies [5, 6, 19, 26], we applied bicubic interpolation to down-sample HR images with 2×, 4×, and 8× ratios. Gaussian blur was then added to generate realistic LR images. Finally, the downsized images were resized to the original image size using bicubic interpolation. Figure 2 presents sample LR images.

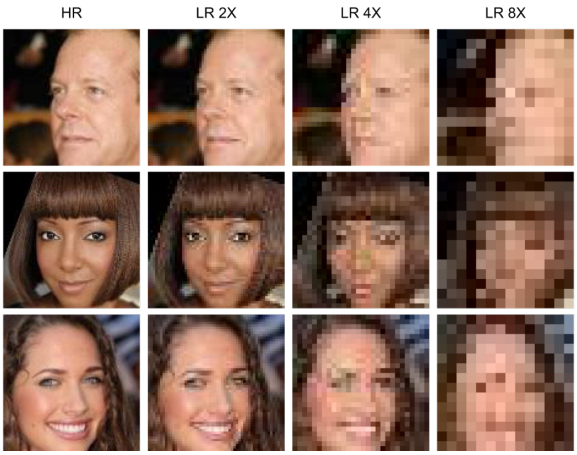


**Fig. 2.** The samples of HR and LR images from the training dataset (CASIA [33]) with the down-sampling ratios of 2×, 4×, and 8×.

### 3.2 Face recognition with attention modules

**Face recognition network.** ArcFace [3] is a SOTA face recognition network comprising convolutional neural network (CNN) backbone and angular margin introduced to softmax loss. Conventional softmax loss can be expressed as

$$L_{softmax} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_j}}, \tag{1}$$

where $x_i \in \mathbb{R}^d$ is the embedded feature of the $i$-th sample belonging to the $y_i$-th class; $N$ and $n$ are the batch size and the number of classes, respectively; $W_j \in \mathbb{R}^d$ denotes the $j$-th column of the last fully connected layer's weight $W \in \mathbb{R}^{d \times n}$ and $b_j \in \mathbb{R}^n$ is the bias term for the $j$-th class.

For simplicity, the bias term is fixed to 0 as in [20]. Then the logit of the $j$-th class can be represented as $W_j^T x_i = \|W_j\|\|x_i\|cos(\theta_j)$, where $\theta_j$ denotes the angle between the $W_j$ and $x_i$. Following previous approaches [20, 29], ArcFace set $\|W_j\| = 1$ and $\|x_i\| = 1$ via $l_2$ normalisation to maximize $\theta_j$ among inter-class and minimize $\theta_j$ among intra-class samples. Further, constant linear angular margin ($m$) was introduced to avoid convergence difficulty. The ArcFace [3] loss can be expressed as

$$L_{arcface} = -\frac{1}{N}\sum_{i=1}^{N} log\frac{e^{s(cos(\theta_{y_i}+m))}}{e^{s(cos(\theta_{y_i}+m))}+\sum_{j=1,j\neq y_i}^{n} e^{s(cos(\theta_j))}}, \tag{2}$$

where $s$ is the re-scale factor and $m$ is the additive angular margin penalty between $x_i$ and $W_{y_i}$.

**Attention.** Attention is a simple and effective method to guide feature focus on important regions for recognition. Let $\mathbf{f}_i = \mathcal{H}_i(\mathbf{x})$ be intermediate feature outputs from the $i$-th layer of the CNN. Attention maps about $\mathbf{f}_i$ can be represented as the $\mathcal{A}_i(\mathbf{f}_i)$, where $\mathcal{A}_i(\cdot)$ is attention module.

Many attention mechanisms have been proposed; AT [35] simply applied channel-wise pooling to features to estimate spatial attention maps. SENet [13] and CBAM [31] utilized parametric transformations, e.g. convolution layers, to represent attention maps. Estimated attention maps were multiplied with the features and passed to a successive layer. Trainable parameters in attention module are updated to improve performance during back-propagation, forming accurate attention maps. Attention mechanisms can be expressed as

$$\mathbf{f_i^{'}} = \mathcal{A}_i^c(\mathbf{f_i}) \otimes \mathbf{f_i} \tag{3}$$

and

$$\mathbf{f_i^{''}} = \mathcal{A}_i^s(\mathbf{f_i^{'}}) \otimes \mathbf{f_i^{'}}, \tag{4}$$

where $\mathcal{A}_i^c(\cdot)$ and $\mathcal{A}_i^s(\cdot)$ are attention modules for channel and spatial attention maps, respectively.

Features are refined twice by multiplying channel and spatial attention maps in order (3) and (4). Any parametric attention transformation could be employed

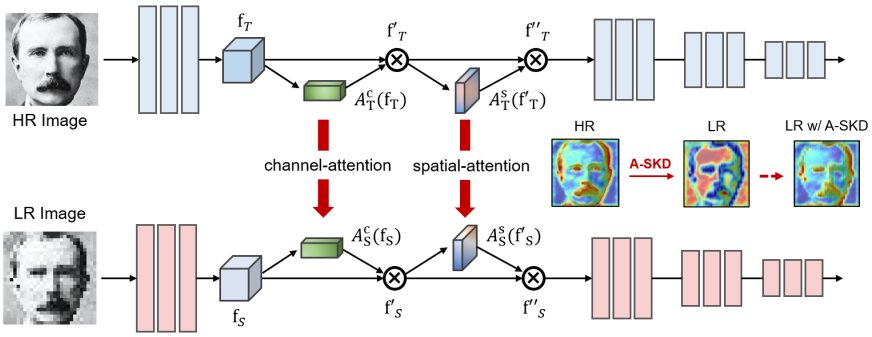


**Fig. 3.** Proposed A-SKD framework. The LR network formulates precise attention maps by referencing well-constructed channel and spatial attention maps obtained from the HR network, focusing on detailed facial parts which are helpful for the face recognition. We only show the attention distillation for the first block.

for the proposed A-SKD, and we adopted the popular CBAM [31] module,

$$\mathcal{A}^c(\mathbf{f}) = \sigma(FC(AvgPool(\mathbf{f})) + FC(MaxPool(\mathbf{f}))) \tag{5}$$

and

$$\mathcal{A}^s(\mathbf{f}) = \sigma(f^{7\times7}(AvgPool(\mathbf{f}); MaxPool(\mathbf{f}))), \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function; and $FC(\cdot)$ and $f^{7\times7}(\cdot)$ are fully connected and convolution layers with $7 \times 7$ filters, respectively.

### 3.3 Proposed attention similarity knowledge distillation framework

Unlike the conventional knowledge distillation, the network size of teacher and student network is same for A-SKD. Instead, the teacher network is trained on HR images whereas the student network is trained on LR images. Due to the resolution differences, features from both networks are difficult to be identical. Therefore, we propose to distill well-constructed attention maps from the HR network into the LR network instead of features.

$$\begin{aligned}\rho_i &= 1 - \langle \mathcal{A}_{T,i}(\mathbf{f}_{T,i}), \mathcal{A}_{S,i}(\mathbf{f}_{S,i}) \rangle \\ &= 1 - \frac{\mathcal{A}_{T,i}(\mathbf{f}_{T,i})}{\|\mathcal{A}_{T,i}(\mathbf{f}_{T,i})\|_2} \cdot \frac{\mathcal{A}_{S,i}(\mathbf{f}_{S,i})}{\|\mathcal{A}_{S,i}(\mathbf{f}_{S,i})\|_2},\end{aligned} \tag{7}$$

where $\rho_i$ is the cosine distance between attention maps from the $i$-th layer of the teacher and student networks; $\langle \cdot, \cdot \rangle$ denotes the cosine similarity; $\|\cdot\|_2$ denotes L2-norm; $\mathcal{A}_i(\mathbf{f}_i)$ denotes the attention maps for the $i$-th layer features; and $T$ and $S$ denote the teacher and student network, respectively. Thus, $\mathcal{A}_{T,i}(\mathbf{f}_{T,i})$ and $\mathcal{A}_{S,i}(\mathbf{f}_{S,i})$ are attention maps estimated from the $i$-th layer of the teacher and

student network's features, respectively. Reducing the cosine distance between HR and LR attention maps increases the similarity between them.

Distillation loss for A-SKD is calculated as

$$\mathcal{L}_{distill} = \sum_{i=1}^{N} \frac{(\rho_i^s + \rho_i^c)}{2} \tag{8}$$

which average the cosine distance for channel and spatial attention maps between the HR and LR networks, and sums them across layers ($i = 1, 2, 3, ..., N$) of the backbone. $N$ is the number of layers utilized for the distillation.

Total loss for the LR face recognition network is the sum of target task's loss and distillation loss (8) weighted by the factor ($\lambda_{distill}$). In this work, we utilized the ArcFace loss (2) as a target task's loss.

$$\mathcal{L}_{total} = \mathcal{L}_{arcface} + \lambda_{distill} * \mathcal{L}_{distill}. \tag{9}$$

Further, our method can be utilized in conjunction with the logit distillation by simply adding the logit distillation loss [12] to our loss function (9). Since logit is the final output of the network, incorporating the logit distillation loss allows the LR network to make the same decision as the HR network based on the refined attention maps.

## 4 Experiments

### 4.1 Settings

**Datasets.** We employed the CASIA [33] dataset for training, which is a large face recognition benchmark comprising approximately 0.5M face images for 10K identities. Each sample in CASIA was down-sampled to construct the HR-LR paired face dataset. For the evaluation, the manually down-sampled face recognition benchmark (AgeDB-30 [23]) and the popular LR face recognition benchmark (TinyFace [1]) were employed. Since AgeDB-30 have similar resolution to CASIA, networks trained on down-sampled CASIA images were validated on AgeDB-30 down-sampled images with matching ratio. In contrast, the real-world LR benchmark (TinyFace) comprises face images with the resolution of 24×24 in average when they are aligned. Therefore, they were validated using a network trained on CASIA images down-sampled to 24×24 pixels.

**Task and metrics.** Face recognition was performed for two scenarios: face verification and identification. Face verification is where the network determines whether paired images are for the same person, i.e., 1:1 comparison. To evaluate verification performance, accuracy was determined using validation sets constructed from probe and gallery set pairs following the LFW protocol [14]. Face identification is where the network recognize the identity of a probe image by measuring similarity against all gallery images, i.e., 1:N comparison. This study employed the smaller AgeDB-30 dataset for the face verification; and larger TinyFace dataset for the face identification.

**Comparison with other methods.** Typically, the distillation of intermediate representation is performed concurrently with the target task's loss. Previous distillation methods in the experiments utilized the both face recognition and distillation loss, albeit face recognition loss of varying forms. In addition, some feature distillation approach reported their performances with the logit distillation loss. In order to conduct a fair comparison, we re-implemented the prior distillation methods with the same face recognition loss (ArcFace [3]) and without the logit distillation loss. Further, our method requires the parametric attention modules for the distillation. Therefore, we utilized the same backbone network with CBAM attention modules for all methods; we combined the CBAM modules to all convolution layers, with the exception of the stem convolution layer and the convolution layer with a kernel size of 1.

**Implementation details.** We followed the ArcFace protocol for data preprocessing: detecting face regions using the MTCNN [36] face detector, cropping around the face region, and resizing the resultant portion to 112×112 pixel using bilinear interpolation. The backbone network was ResNet-50 with CBAM attention module. For the main experiments, we distilled the attention maps for every convolution layers with the exception of the stem convolution layer and the convolution layer with a kernel size of 1. Weight factors for distillation (9) $\lambda_{distill} = 5$. This weight factors generally achieved the superior results not only for the face recognition benchmarks, but also for the ImageNet [2]. Learning rate = 0.1 initially, divided by 10 at 6, 11, 15, and 17 epochs. SGD optimizer was utilized for the training with batch size = 128. Training completed after 20 epochs. The baseline refers to the LR network that has not been subjected to any knowledge distillation methods. For the hyperparameter search, we divided 20% of the training set into the validation set and conducted a random search. After the hyperparameter search, we trained the network using the whole training set and and performed the evaluation on the AgeDB-30 and TinyFace.

### 4.2 Face recognition benchmark results

**Evaluation on AgeDB-30.** Table 1 shows LR face recognition performance on AgeDB-30 with various down-sample ratios depending on distillation methods. Except for HORKD, previous distillation methods [22, 35] exhibited only slight improvement or even reduced performance when the downsampling ratios increase. This indicates that reducing the L2 distance between the HR and LR network's features is ineffective. In contrast, HORKD improved LR recognition performance by distilling the relational knowledge of the HR network's features. When the input's resolution decrease, the intermediate features are hard to be identical with the features from the HR network. Instead the relation among the features of the HR network can be transferred to the LR network despite the spatial information loss; this was the reason of HORKD's superior performances even for the 4× and 8× settings.

However, attention maps from the HORKD exhibit similar pattern to LR baseline network rather than the HR network in the Figure 4. HR attention maps are highly activated in facial landmarks, such as eyes, lips, and beard, which

**Table 1.** Proposed A-SKD approach compared with baseline and previous SOTA methods on AgeDB-30 with 2×, 4×, and 8× down-sampled ratios. L, F, SA, and CA indicate distillation types of logit, feature, spatial attention, and channel attention, respectively. Ver-ACC denotes the verification accuracy. Base refers to the LR network that has not been subjected to any knowledge distillation methods.

| Resolution | Method | Distill Type | Loss Function | Ver-ACC (%) (AgeDB-30) |
|---|---|---|---|---|
| 1× | Base | - | - | 93.78 |
| 2× | Base | - | - | 92.83 |
| | F-KD [22] | F | L2 | 93.05 |
| | AT [35] | SA | L2 | 92.93 |
| | HORKD [8] | F | L1+Huber | 93.13 |
| | A-SKD **(Ours)** | SA+CA | Cosine | **93.35** |
| | A-SKD+KD **(Ours)** | SA+CA+L | Cosine+KLdiv | **93.58** |
| 4× | Base | - | - | 87.74 |
| | F-KD [22] | F | L2 | 87.72 |
| | AT [35] | SA | L2 | 87.75 |
| | HORKD [8] | F | L1+Huber | 88.08 |
| | A-SKD **(Ours)** | SA+CA | Cosine | **88.58** |
| | A-SKD+KD **(Ours)** | SA+CA+L | Cosine+KLdiv | **89.15** |
| 8× | Base | - | - | 77.75 |
| | F-KD [22] | F | L2 | 77.85 |
| | AT [35] | SA | L2 | 77.40 |
| | HORKD [8] | F | L1+Huber | 78.27 |
| | A-SKD **(Ours)** | SA+CA | Cosine | **79.00** |
| | A-SKD+KD **(Ours)** | SA+CA+L | Cosine+KLdiv | **79.45** |

are helpful features for face recognition [18]. In contrast, detailed facial parts are less activated for LR attention maps because those parts are represented with a few pixels. Although HORKD boosts LR recognition performance by transferring HR relational knowledge, it still failed to capture detailed facial features crucial for recognition. The proposed A-SKD method directs the LR network's attention toward detailed facial parts that are well represented by the HR network's attention maps.

Based on the refined attention maps, A-SKD outperforms the HORKD and other knowledge distillation methods for all cases. AgeDB-30 verification accuracy increased 0.6%, 1.0%, and 1.6% compared with baseline for 2×, 4×, and 8× down-resolution ratios, respectively. In addition, when A-SKD is combined with logit distillation (KD), the verification accuracy increased significantly for all settings. From the results, we confirmed that the attention knowledge from the HR network can be transferred to the LR network and led to significant improvements that were superior to the previous SOTA method.

**Evaluation on TinyFace.** Unlike the face verification, the identification task requires to select a target person's image from the gallery set consists of a large number of face images. Therefore, the identification performances decrease significantly when the resolution of face images are degraded. Table 2 showed the identification performances on the TinyFace benchmark. When the AT [35] was applied, the rank-1 identification accuracy decreased 13.34% compared to the

baseline. However, our approach improved the rank-1 accuracy 13.56% compared to the baseline, even outperforming the HORKD method. This demonstrated that the parametric attention modules (CBAM) and cosine similarity loss are the key factors for transferring the HR network's knowledge into the LR network via attention maps. The proposed method is generalized well to real-world LR face identification task which is not manually down-sampled.

**Table 2.** Evaluation results on TinyFace identification benchmark depending on the distillation methods. Acc@K denotes the rank-K accuracy (%).

| | **ACC@1** | **ACC@5** | **ACC@10** | **ACC@50** |
|---|---|---|---|---|
| Base | 42.19 | 50.62 | 53.67 | 60.41 |
| AT [35] | 36.56 | 45.68 | 49.03 | 56.44 |
| HORKD [8] | 45.49 | 54.80 | 58.26 | 64.30 |
| A-SKD **(Ours)** | **47.91** | **56.55** | **59.92** | **66.60** |

## 5 Discussion

**Attention correlation analysis.** Figure 5 shows Pearsons correlation between attention maps from the HR and LR networks for the different distillation methods. Spatial and channel attention maps from the four blocks for models other than A-SKD have a low correlation between the HR and LR networks, with a magnitude lower than 0.5. In particular, spatial attention maps obtained from the first block of the LR baseline and HORKD network have negative correlation with the HR network ($r = -0.39$ and $-0.29$, respectively).

Figure 4 shows that spatial attention maps from LR baseline and HORKD networks are highly activated in skin regions, which are less influenced by resolution degradation, in contrast to the HR network. This guides the LR network to the opposite directions from the HR network. However, spatial attention maps from A-SKD exhibit strong positive correlation with those from the HR network, highlighting detailed facial attributes such as beard, hair, and eyes. Through the A-SKD, the LR network learned where to focus by generating precise attention maps similar to those for the HR network. Consequently, Pearsons correlation, i.e., the similarity measure between HR and LR attention maps, was significantly improved for all blocks, with a magnitude higher than 0.6. Thus the proposed A-SKD approach achieved superior efficacy and success compared with previous feature based SOTA methods.

**Comparison with attention transfer [35].** Primary distinctions between AT [35] and A-SKD include the cosine similarity loss, parametric attention modules, and distillation of both channel and spatial attention maps. Correlation analysis for A-SKD confirmed that the cosine similarity loss is an effective strategy for transferring attention knowledge. Distilling AT attention maps using the


**Fig. 4.** Normalized spatial attention maps from the first block for different distillation methods. Red and blue regions indicate high and low attention, respectively. Face images and attention maps are from the AgeDB-30.

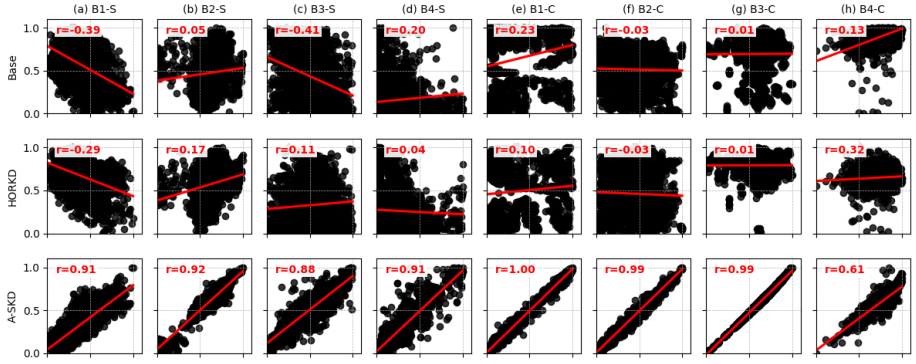


**Fig. 5.** Pixel level Pearsons correlation between the HR and LR network's attention maps for different distillation methods. B{$i$}-{S,C} indicates Pearsons correlation for spatial or channel attention maps obtained from the $i$-th ResNet block between the HR and LR networks; and $r$ is Pearsons correlation coefficient representing linear relationships between input variables. Base refers to the LR network that has not been subjected to any knowledge distillation methods. Pearsons correlation is measured using the AgeDB-30.

cosine similarity rather than the L2 loss increased AgeDB-30 verification accuracy by 0.32%p (Table 3). AT calculates attention maps using channel-wise pooling, a non-parametric layer; whereas A-SKD calculates attention maps using parametric layers comprising fully connected and convolution layers. When the input image resolution degrades, the student network's feature representation diverges from that of the teacher network. Therefore, it is difficult to match the attention maps of the student network obtained by the non-parametric module with those of the teacher network. Instead, A-SKD employs the parametric module for the attention maps extraction and the cosine similarity loss for the distillation; therefore, the attention maps from the student network can be adaptively trained to be similar to the attention maps from the teacher network despite the differences in the features. Finally, A-SKD distills both spatial and channel attention maps in contrast to AT which only considered spatial attention maps. We confirmed A-SKD with spatial and channel attention additionally improved AgeDB-30 verification accuracy by 0.34%p compared with spatial-only attention. This comparison results also confirmed that A-SKD, designed for attention distillation on LR settings, is the most effective approach for transferring attention knowledge.

**Table 3.** Comparing attention transfer (AT) [35] and proposed A-SKD on AgeDB-30 benchmark down-sampled with 8× ratio. AT* indicates the cosine similarity loss was utilized for attention transfer rather than the original L2 loss. SA and CA indicate spatial and channel attention maps, respectively.

| **Method** | **Type** | **Transformation** | **Loss Function** | **Ver-ACC (%) (AgeDB-30)** |
|---|---|---|---|---|
| AT [35] | SA | Non-parametric layer | L2 | 77.40 |
| AT* | SA | Non-parametric layer | Cosine | 77.72 |
| A-SKD | SA | Parametric layer | Cosine | 78.66 |
| A-SKD | SA + CA | Parametric layer | Cosine | **79.00** |

## 6 Extension to Other Tasks

### 6.1 Object classification

We conducted experiments for object classification on LR images using the 4× down-sampled ImageNet [2]. For the backbone network, we utilized the ResNet18 with CBAM attention modules. We compared our method to other knowledge distillation methods (AT [35] and RKD [25]) which are widely utilized in the classification domains. We re-implemented those methods using its original hyperparameters. Usually, AT and RKD were utilized along with the logit distillation for the ImageNet; therefore, we performed the AT, RKD, and A-SKD in conjunction with the logit distillation in the Table 4. Training details are provided in the Supplementary Information.

Table 4 shows that A-SKD outperformed the other methods on the LR ImageNet classification task. Park et al. demonstrated that introducing the accurate attention maps led the significant improvement on classification performances [24, 31]. When the attention maps were distilled from the teacher network, student network could focus on informative regions by forming precise attention maps similar with the teacher's one. Thus, our method can be generalized to general object classification task, not restricted to face related tasks.

**Table 4.** Proposed A-SKD performance on low resolution ImageNet classification. All distillation methods were performed in conjunction with the logit distillation.

| Resolution | Method | ACC (%) |
|---|---|---|
| 1× | Base | 70.13 |
| 4× | Base | 65.34 |
| | AT [35] | 65.79 |
| | RKD [25] | 65.95 |
| | A-SKD | **66.52** |

### 6.2 Face detection

Face detection is a sub-task of object detection to recognize human faces in an image and estimate their location(s). We utilized TinaFace [38], a deep learning face detection model, integrated with the CBAM attention module to extend the proposed A-SKD approach to face detection. Experiments were conducted on the WIDER FACE [32] dataset (32,203 images containing 393,703 faces captured from real-world environments) with images categorized on face detection difficulty: easy, medium, and hard. LR images were generated with 16× and 32× down-resolution ratios, and further training and distillation details are provided in the Supplementary Information.

**Table 5.** Proposed A-SKD performance on LR face detection. mAP is mean average precision; easy, medium, and hard are pre-assessed detection difficulty.

| Resolution | Model | mAP (%) | | |
|---|---|---|---|---|
| | | Easy | Medium | Hard |
| 1× | Base | 95.56 | 95.07 | 91.45 |
| 16× | Base | 54.38 | 52.73 | 35.29 |
| | A-SKD | **62.93** | **60.19** | **47.28** |
| 32× | Base | 31.15 | 26.68 | 14.00 |
| | A-SKD | **33.50** | **30.04** | **16.02** |

Table 5 shows that A-SKD improved the overall detection performance by distilling well-constructed attention maps, providing significant increases of mean

average precision (mAP) for the easy (15.72% for 16× and 7.54% for 32×), medium (14.15% for 16× and 12.59% for 32×), and hard (33.98% for 16× and 14.43% for 32×) level detection tasks. Small faces were well detected in the LR images after distillation as illustrated in Figure 6. Thus the proposed A-SKD approach can be successfully employed for many LR machine vision tasks.


**Fig. 6.** Qualitative results for LR face detection before and after applying A-SKD. Small faces were better detected after A-SKD. The face images are from the evaluation set of WIDERFACE.

## 7 Conclusion

We verified that attention maps constructed from HR images were simple and effective knowledge that can be transferred to LR recognition networks to compensate for spatial information loss. The proposed A-SKD framework enabled any student network to focus on target regions under LR circumstances and generalized well for various LR machine vision tasks by simply transferring well-constructed HR attention maps. Thus, A-SKD could replace conventional KD methods offering improved simplicity and efficiency and could be widely applicable to LR vision tasks, which have not been strongly studied previously, without being limited to face related tasks.

**Acknowledgments** This work was supported by the ICT R&D program of MSIT/IITP[2020-0-00857, Development of Cloud Robot Intelligence Augmentation, Sharing and Framework Technology to Integrate and Enhance the Intelligence of Multiple Robots. And also, this work was partially supported by Korea Institute of Energy Technology Evaluation and Planning (KETEP) grant funded by the Korea government (MOTIE)(No. 20202910100030) and supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government. [22ZR1100, A Study of Hyper-Connected Thinking Internet Technology by autonomous connecting, controlling and evolving ways].

## References


1. Cheng, Z., Zhu, X., Gong, S.: Low-resolution face recognition. In: ACCV (2018)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009). https://doi.org/10.1109/CVPR.2009.5206848
3. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2019-June**, 4685–4694 (jan 2018), http://arxiv.org/abs/1801.07698
4. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**, 295–307 (2016)
5. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: ECCV (2016)
6. Flusser, J., Farokhi, S., Höschl, C., Suk, T., Zitová, B., Pedone, M.: Recognition of Images Degraded by Gaussian Blur. IEEE Transactions on Image Processing **25**(2), 790–806 (2016). https://doi.org/10.1109/TIP.2015.2512108
7. Fookes, C., Lin, F., Chandran, V., Sridharan, S.: Evaluation of image resolution and super-resolution on face recognition performance. Journal of Visual Communication and Image Representation **23**(1), 75–93 (jan 2012). https://doi.org/10.1016/j.jvcir.2011.06.004
8. Ge, S., Zhang, K., Liu, H., Hua, Y., Zhao, S., Jin, X., Wen, H.: Look One and More: Distilling Hybrid Order Relational Knowledge for Cross-Resolution Image Recognition. Proceedings of the AAAI Conference on Artificial Intelligence **34**(07), 10845–10852 (2020). https://doi.org/10.1609/aaai.v34i07.6715, https://ojs.aaai.org/index.php/AAAI/article/view/6715
9. Gunturk, B.K., Batur, A.U., Altunbasak, Y., Hayes, M.H., Mersereau, R.M.: Eigenface-domain super-resolution for face recognition. IEEE Transactions on Image Processing **12**(5), 597–606 (may 2003). https://doi.org/10.1109/TIP.2003.811513
10. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **9907 LNCS**, 87–102 (jul 2016), http://arxiv.org/abs/1607.08221
11. Hennings-Yeomans, P.H., Baker, S., Kumar, B.V.: Simultaneous super-resolution and feature extraction for recognition of low-resolution faces. In: 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR (2008). https://doi.org/10.1109/CVPR.2008.4587810
12. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015), http://arxiv.org/abs/1503.02531
13. Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E.: Squeeze-and-excitation networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **42**, 2011–2023 (2020)
14. Huang, G.B., Mattar, M.A., Berg, T.L., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)

15. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4873–4882 (2016). https://doi.org/10.1109/CVPR.2016.527
16. Kong, H., Zhao, J., Tu, X., Xing, J., Shen, S., Feng, J.: Cross-Resolution Face Recognition via Prior-Aided Face Hallucination and Residual Knowledge Distillation. arXiv (may 2019), http://arxiv.org/abs/1905.10777
17. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops) pp. 2144–2151 (2011)
18. Kumar, A., Chellappa, R.: S2ld: Semi-supervised landmark detection in low resolution images and impact on face verification. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) pp. 3275–3283 (2020)
19. Ledig, C., Theis, L., Huszár, F., Caballero, J., Aitken, A., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 105–114 (2017)
20. Liu, W., Wen, Y., Yu, Z., Li, M., Raj, B., Song, L.: Sphereface: Deep hypersphere embedding for face recognition. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 6738–6746 (2017)
21. Lui, Y.M., Bolme, D., Draper, B.A., Beveridge, J.R., Givens, G., Phillips, P.J.: A Meta-Analysis of Face Recognition Covariates. In: Proceedings of the 3rd IEEE International Conference on Biometrics: Theory, Applications and Systems. pp. 139–146. BTAS'09, IEEE Press (2009)
22. Massoli, F.V., Amato, G., Falchi, F.: Cross-resolution learning for Face Recognition. Image and Vision Computing **99**, 103927 (jul 2020). https://doi.org/10.1016/j.imavis.2020.103927
23. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: AgeDB: The First Manually Collected, In-the-Wild Age Database. pp. 1997–2005 (2017). https://doi.org/10.1109/CVPRW.2017.250
24. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: Bottleneck attention module. In: BMVC (2018)
25. Park, W., Kim, D., Lu, Y., Cho, M.: Relational Knowledge Distillation. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition **2019-June**, 3962–3971 (apr 2019), http://arxiv.org/abs/1904.05068
26. Pei, Y., Huang, Y., Zou, Q., Zhang, X., Wang, S.: Effects of image degradation and degradation removal to cnn-based image classification. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(4), 1239–1253 (2021). https://doi.org/10.1109/TPAMI.2019.2950923
27. Romero, A., Ballas, N., Kahou, S.E., Chassang, A., Gatta, C., Bengio, Y.: Fitnets: Hints for thin deep nets. CoRR **abs/1412.6550** (2015)
28. Tran, L., Yin, X., Liu, X.: Disentangled representation learning GAN for pose-invariant face recognition. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017. vol. 2017-January, pp. 1283–1292. Institute of Electrical and Electronics Engineers Inc. (nov 2017). https://doi.org/10.1109/CVPR.2017.141
29. Wang, H., Wang, Y., Zhou, Z., Ji, X., Li, Z., Gong, D., Zhou, J., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 5265–5274 (2018)

30. Wilman, W.W.Z., Yuen, P.C.: Very low resolution face recognition problem. In: 2010 Fourth IEEE International Conference on Biometrics: Theory, Applications and Systems (BTAS). pp. 1–6 (2010). https://doi.org/10.1109/BTAS.2010.5634490
31. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: CBAM: Convolutional Block Attention Module. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) **11211 LNCS**, 3–19 (jul 2018), http://arxiv.org/abs/1807.06521
32. Yang, S., Luo, P., Loy, C.C., Tang, X.: WIDER FACE: A Face Detection Benchmark. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5525–5533 (2016). https://doi.org/10.1109/CVPR.2016.596
33. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning Face Representation from Scratch (nov 2014), http://arxiv.org/abs/1411.7923
34. Yim, J., Joo, D., Bae, J.H., Kim, J.: A gift from knowledge distillation: Fast optimization, network minimization and transfer learning. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) pp. 7130–7138 (2017)
35. Zagoruyko, S., Komodakis, N.: Paying More Attention to Attention: Improving the Performance of Convolutional Neural Networks via Attention Transfer. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings (dec 2016)
36. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks. IEEE Signal Processing Letters **23**(10), 1499–1503 (apr 2016). https://doi.org/10.1109/LSP.2016.2603342, http://arxiv.org/abs/1604.02878 http://dx.doi.org/10.1109/LSP.2016.2603342
37. Zhu, M., Han, K., Zhang, C., Lin, J., Wang, Y.: Low-resolution Visual Recognition via Deep Feature Distillation. In: ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3762–3766 (2019). https://doi.org/10.1109/ICASSP.2019.8682926
38. Zhu, Y., Cai, H., Zhang, S., Wang, C., Xiong, Y.: Tinaface: Strong but simple baseline for face detection. ArXiv **abs/2011.13183** (2020)