https://doi.org/10.2352/EI.2026.38.6.ISS-278

FlatFace: Improve Face Recognition by Sharpness-Aware Minimization

Tanapat Ratchatorn and Masayuki Tanaka;
Institute of Science Tokyo, Tokyo, Japan

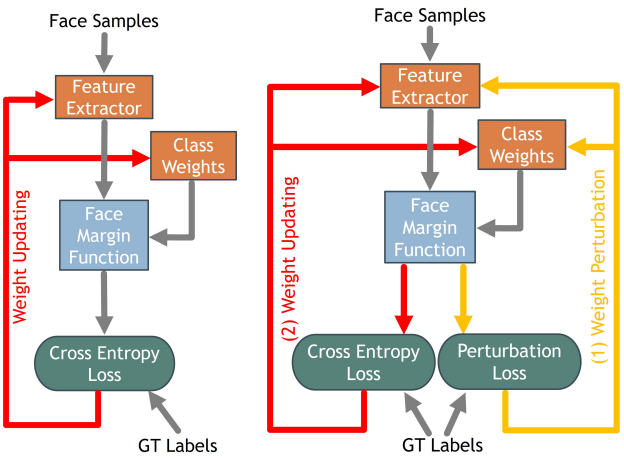
Abstract

Margin-based Face Recognition (FR) has achieved remarkable performance by learning discriminative feature representations that ensure high intra-class compactness and inter-class separability. While most state-of-the-art methods focus on developing margin-based loss functions, improving model generalization performance is equally critical, especially under open-set conditions where test identities are absent from training data. Recent developments in learning algorithms have highlighted the sharpness of the loss surface as a key factor in reducing the generalization gap. Building on this, Sharpness-Aware Minimization (SAM) introduced a weight perturbation step to enhance generalization performance, with Adaptive Adversarial Cross-Entropy (AACE) further refining SAM by modifying the perturbation step. Inspired by those researches, we propose FlatFace, a novel training framework for face recognition that adopts weight perturbation into the training process. FlatFace consists of two key steps: the perturbation step, which perturbs model parameters in both the feature extractor and class weights toward the worst-case scenario, and the weight updating step, which uses the loss gradient at the perturbed feature extractor and class weights to update the parameters. By guiding the model toward flatter minima, FlatFace improves generalization performance and accuracy, particularly for open-set face recognition tasks. Empirical experiments confirm its effectiveness, demonstrating reduced generalization gaps and enhanced overall performance.

Introduction

In recent years, Face Recognition (FR) has made extraordinary strides due to advancements in Deep Convolutional Neural Networks (DCNNs) and large-scale datasets [1, 2, 3, 4]. The pivotal challenge in FR lies in learning discriminative feature representations that ensure high intra-class compactness and inter-class separability, particularly under open-set conditions where identities in testing data are unseen during training [5, 6, 7].

Traditional softmax loss functions, though widely used, still lack the discriminative power required for open-set face recognition. To address this limitation, several research studies that modify loss function have been proposed [8, 9, 10]. Among the numerous modified loss functions, margin-based loss functions have been introduced and have demonstrated remarkable performance. SphereFace introduces the concept of multiplicative angular margin by enforcing feature separability on a hypersphere manifold, demonstrating significant improvements in intra-class compactness [11]. CosFace refined this concept by employing an additive cosine margin penalty for easier optimization [12]. ArcFace further advanced this field with additive angular margin loss, directly optimizing feature angular discrimination [13]. CurricularFace incorporated curriculum learning to dynamically adjust margin constraints[14]. AdaFace proposed adaptive feature normalization based on individual sample qualities, improving robustness for real-world applications [15].



(a) Existing margin-based FR's training procedure.

(b) Proposed FlatFace's training procedure.

Figure 1: Training process comparison between normal margin-based FR and FlatFace.

While those margin-based loss functions can enhance face recognition performance, they are not the sole solution, especially for open-set tasks. Improving model generalization is equally important to ensure effective performance on unseen data, a crucial requirement for open-set conditions. Among the various aspects of research related to model generalization, some researchers explored the relationship between the shape of the loss landscape and model generalization. Findings indicate that reducing the sharpness of the loss surface and optimizing generalization bounds are critical for achieving superior performance across various tasks [16, 17, 18].

Sharpness-Aware Minimization (SAM) has been proposed to find flatter regions in the loss landscape by introducing small perturbations to model parameters and has proven to be both versatile and effective across diverse tasks [19]. In addition to SAM, the Adaptive Adversarial Cross-Entropy (AACE) Loss has been proposed to replace the standard cross-entropy loss and modify the perturbation calculation in SAM's perturbation step, demonstrating further enhancements [20].

In this work, we propose a new training framework named FlatFace for the FR. Inspired by SAM [19] and AACE [20], we adopt the concept of parameter perturbation into the margin-based

FR. Similar to general SAM's framework, our proposed framework consists of two main steps, as shown in Fig. 1 (b). First, the perturbation vectors are obtained using the gradients of the loss at the current position. These calculated perturbation vectors are used to perturb the learnable parameters, in both the feature extractor and the class weight vector, toward the worst-case scenario. Then, the model weights are updated using the new gradients calculated at the perturbed position. Moreover, we also suggest to integrate AACE loss into the perturbation step which helps enhance the perturbation stability further.

Since FlatFace is a training framework designed as an extension to FR training, it can be used with various margin-based FR methods and helps encourages the model to the flatter region in the loss landscape which leads to better generalization and overall robustness of the open-set face recognition task.

Several experiments were conducted using different FR techniques, with ArcFace and AdaFace chosen as they are among the most popular state-of-the-art methods. The results on various benchmarks highlight the possibility of applying this new training framework to different FR methods.

Releated Works

Margin-Based Face Recognition

A standard softmax loss, which is a softmax function followed by a standard cross-entropy loss computation, of a set of samples with batch size N can be expressed as:

$$L = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{u_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{C} e^{u_j^T x_i + b_j}}, \tag{1}$$

where $x_i \in \mathbb{R}^d$ is the feature vector of the i-th sample, $u_j \in \mathbb{R}^d$ is the class j-th weight vector, b_j is the bias term, y_i denotes the ground-truth class, and C is the number of classes.

For simplicity, the bias is usually set to $b_j = 0$, while we also fix $\|u_j\| = 1$ by l_2 norm, and the feature is normalized and re-scaled by s. Moreover, since $\cos(\theta_j) = \frac{u_j^T x_i}{\|u_j\|\|x_i\|}$ is the cosine similarity between x_i and u_j, the softmax loss can now be re-written as:

$$L_{\text{normalized}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i})}}{\sum_{j=1}^{C} e^{s\cos(\theta_j)}}. \tag{2}$$

The softmax loss lacks explicit constraints on intra-class compactness and inter-class separability, limiting its discriminative power.

Margin-based face recognition has emerged as a significant approach to improving feature discriminability in deep learning-based facial recognition systems. The key idea is to enforce larger inter-class margins and compact intra-class distributions in the embedding space, leading to enhanced separation between different facial identities. Several modifications have been proposed to increase the discriminative power of softmax loss.

One of the most widely used margin-based FR methods is ArcFace [13], which introduces an additive angular margin loss, which directly optimizes the geodesic distance on a normalized hypersphere. This is achieved by modifying the standard softmax loss to include an angular margin m in the computation. The formulation is given by:

$$L_{\text{ArcFace}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos(\theta_{y_i}+m)}}{e^{s\cos(\theta_{y_i}+m)} + \sum_{j\neq y_i}^{C} e^{s\cos(\theta_j)}}. \tag{3}$$

Another example is AdaFace [15], which further advances margin-based face recognition by adapting the margin dynamically based on the quality of the input image. This is achieved by using the feature norm as a proxy for image quality, allowing the loss to emphasize the harder yet recognizable samples while de-emphasizing the samples with low quality. The formula can be expressed as:

$$L_{\text{AdaFace}} = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+g_{\text{angle}})-g_{\text{add}})}}{e^{s(\cos(\theta_{y_i}+g_{\text{angle}})-g_{\text{add}})} + \sum_{j\neq y_i}^{C} e^{s\cos(\theta_j)}}, \tag{4}$$

where

$$g_{\text{angle}} = -m\cdot\widehat{\|x_i\|}, \quad g_{\text{add}} = m\cdot\widehat{\|x_i\|} + m, \tag{5}$$

and the normalized feature norm, $\widehat{\|x_i\|}$, can be computed from:

$$\widehat{\|x_i\|} = \left\lfloor \frac{\|x_i\| - \mu_x}{\sigma_x/h} \right\rceil_{-1}^{1}, \tag{6}$$

where μ_x and σ_x are mean and standard deviation of all feature norms $\|x_i\|$, $\lfloor\cdot\rceil$ denotes clipping the value between -1 and 1, and h is a pre-defined value controlling the concentration.

Sharpness-Aware Minimization and Adaptive Adversarial Cross-Entropy Loss

Traditional deep neural network training methods, aim to minimize the loss function. However, this process often results in convergence to sharp minima in the parameter space where the loss is low for training data but can be high for unseen data. These sharp minima are considered less robust and tend to generalize poorly compared to flat minima.

Stochastic Gradient Descent (SGD) [21] optimizes model parameters by:

$$w_{t+1}^{\text{SGD}} = w_t - \eta\nabla L(w_t), \tag{7}$$

where w_t represents the model parameters at iteration t, $L(w_t)$ is the training loss, and η is the learning rate. The SGD frequently settles at sharp minima. Sharpness-Aware Minimization (SAM) [19] introduces an innovative training approach to improve the generalization performance of deep learning models. SAM's optimization aims to find parameters that minimize the loss not only at the current position but also within a surrounding neighborhood:

$$w_{t+1}^{\text{SAM}} = w_t - \eta\nabla L(w_t + \varepsilon), \tag{8}$$

where

$$\varepsilon = \text{StopGrad}\left(\rho\frac{\nabla L(w_t)}{\|\nabla L(w_t)\|_2}\right) \tag{9}$$

is a perturbation vector that identifies the direction in the parameter space where the loss increases most steeply, scaled by the neighborhood radius ρ. The StopGrad denotes that ε is solely used for the perturbation and remains fixed during the gradient computation for weight updates. This method guides the optimizer toward flatter minima, which are generally considered to offer better generalization to unseen data.

SAM conventionally uses cross-entropy loss for both the perturbation and update steps. However, this is not a strict requirement, as SAM can be applied with any type of loss function. Hence, Adaptive Adversarial Cross-Entropy (AACE) loss, which increases in both value and gradient magnitude as the model converges, was proposed to replace the standard cross-entropy loss in SAM's perturbation calculation, ensuring more stable perturbation directions. Moreover, to make growing perturbation, which is more preferred than fixed radius perturbation, AACE loss has been proposed to modify the perturbation vector by not normalizing the gradient [20]. As a result, Eq. 9 can be replaced with:

$$\varepsilon_{\text{AACE}} = -\text{StopGrad}(\rho \nabla L_{\text{AACE}}(w_t)), \tag{10}$$

where $L_{\text{AACE}}(w_t)$ is an AACE loss which can be defined as:

$$L_{\text{AACE}}(w_t) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j}\tau_j^{\text{AACE}}\log(q_j). \tag{11}$$

Here, q_j is the predicted probability corresponding to class j, and τ_j^{AACE} is AACE loss's target probability distribution for class j:

$$\tau_j^{\text{AACE}} = \xi(\tilde{q}_j), \tag{12}$$

where

$$\xi(\tilde{q}_j) = \begin{cases} 0, & (j = y_i) \\ \frac{\tilde{q}_j}{\sum_{j \neq y_i}\tilde{q}_j}, & (j \neq y_i) \end{cases}, \tag{13}$$

and

$$\tilde{q}_j = \text{StopGrad}(q_j). \tag{14}$$

In this paper, we call this method AACE in short. AACE helps enhance SAM's stability and further improves SAM's performance.

Proposed Method

For open-set face recognition, where real-world applications primarily focus on unseen test data, enhancing model generalization is crucial. It ensures effective performance not only on the training data but also on unseen data, which is a vital requirement for this task.

In this study, we draw inspiration from SAM. Specifically, we propose a new training framework to train the margin-based FR by incorporating the parameters perturbation step into the training loop. This method is called FlatFace.

Fig. 1 (a) represents the overall training process of the conventional FR methods. From an optimizer perspective, the weight updating formula is:

$$w_{t+1}^{\text{FR}} = w_t - \eta \nabla L_{\text{margin}}(w_t), \tag{15}$$

where w_t represents the model weights, including all the learnable parameters in the feature extractor and the classification weights. $L_{\text{margin}}(w_t)$ represents margin-based loss function, such as L_{ArcFace} or L_{AdaFace}. Here, $L_{\text{margin}}(w_t)$ of the sample (x, y) with model weights w_t can be expressed as:

$$L_{\text{margin}}(w_t) = L_{CE}(f(x,u), y), \tag{16}$$

where $f(x,u)$ indicates the margin-based modified logits. In practice, $L_{\text{margin}}(w_t)$ is just a cross-entropy loss of the output from the margin function. Also, SAM can work with various types of loss functions, it is possible to adopt a framework that modifies the training process of margin-based FR by simply incorporating a weight perturbation step, without requiring further modifications.

In our framework, similar to conventional SAM, the overall algorithm can be considered as two main steps. First, the perturbation step, the algorithm performs backpropagation to obtain the loss's gradients needed for the perturbation vector calculation, then it shifts the model configuration to the worst-case scenario within a small neighborhood around the current weights where the loss is maximized. The perturbation vector used for this shift can be calculated from:

$$\varepsilon_{\text{FlatFace-SAM}} = \text{StopGrad}\left(\rho \frac{\nabla L_{\text{margin}}(w_t)}{\left\|\nabla L_{\text{margin}}(w_t)\right\|_2}\right). \tag{17}$$

Then, for the second step, or the weight updating step, it updates the model parameters at the original position by optimizing the parameters with the gradients of the loss computed at the perturbed position $(w_t + \varepsilon_{\text{FlatFace-SAM}})$. As a result, the parameters updating formula can be summarized as:

$$w_{t+1}^{\text{FlatFace-SAM}} = w_t - \eta \nabla L_{\text{margin}}(w_t + \varepsilon_{\text{FlatFace-SAM}}). \tag{18}$$

This variation of FlatFace is called FlatFace-SAM.

Additionally, since AACE loss has been proven to stabilize SAM's perturbation, and often improves generalization [20], we also suggest to use AACE's perturbation. This modification is named FlatFace-AACE, which can be expressed as:

$$w_{t+1}^{\text{FlatFace-AACE}} = w_t - \eta \nabla L_{\text{margin}}(w_t + \varepsilon_{\text{FlatFace-AACE}}), \tag{19}$$

where

$$\varepsilon_{\text{FlatFace-AACE}} = -\text{StopGrad}(\rho \nabla L_{\text{margin}}^{\text{AACE}}(w_t)), \tag{20}$$

and

$$L_{\text{margin}}^{\text{AACE}}(w_t) = L_{\text{AACE}}(f(x,u), y), \tag{21}$$

which is basically an AACE loss of the output from the margin function.

In this paper, we use the term FlatFace-[Perturbation Method]-[FR method name] to refer to the setup. For example, FlatFace-SAM-ArcFace indicates ArcFace trained with FlatFace framework using standard SAM's perturbation and FlatFace-AACE-AdaFace represents AdaFace-based training with AACE's perturbation.

Since this training framework is designed to be used as an extension to the FR training process, it is compatible with most FR methods. We believe this framework can guide the optimizer toward flatter minima, which leads to superior generalization and better performance on unseen data.

Experiments

Implementation Details

For the implementation, we trained the models on the CASIA-WebFace dataset [22] and MS1MV2 [13] with ResNet50

Table 1: The best ρ value for each setup.

setup	FlatFace-SAM-		FlatFace-AACE-	
	ArcFace	AdaFace	ArcFace	AdaFace
ρ	0.1	0.05	0.05	0.2

and ResNet100 as backbones. The dataset preparations followed [13, 14, 15], the samples were cropped to 112 x 112 with five landmarks. We used SGD as a base optimizer in SAM's algorithm with momentum 0.9 and weight decay 5e-4. For CASIA-Webface, the models were trained for 50 epochs with an initial learning rate of 0.1 and divided by 10 at epochs 28, 38, and 46. On MS1MV2, we trained the models for 24 epochs and divided the learning rate at epochs 10, 18, and 22. ArcFace and AdaFace, two widely-used state-of-the-art methods, were chosen as base FR methods for the experiments. We set the scale parameter for margin functions to 64. As in the original works, for ArcFace-based experiments, m was set to 0.5, and for AdaFace-based, we set m to 0.4 and h to 0.33. As for the testing, we used LFW [23], CFP-FP [24], AgeDB [25], CPLFW [26], and CALFW [27] datasets as our benchmarks.

Hyperparameters Tuning

Applying the parameters perturbation step into the algorithm requires an additional parameter, neighborhood radius ρ. We realized that increased sensitivity to hyperparameter selection can add complexity to the practical application. Therefore, we first conducted the experiments on various setups, using different ρ to find the optimum hyperparameter for each setup. We then kept using the obtained ρ for the rest of the experiments in this paper to demonstrate that our methods can be utilized without requiring re-tuning of this hyperparameter for every single setup. The best ρ values for each setup are shown in Table 1.

Results on CASIA-Webface

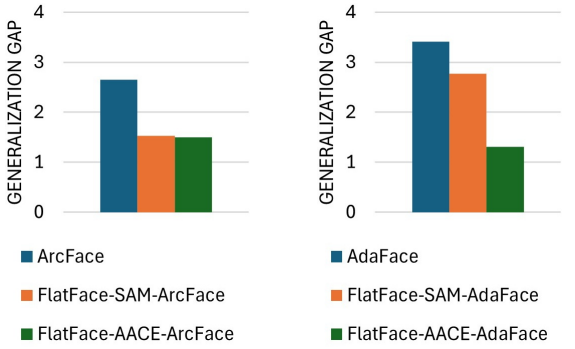
The empirical experiments were conducted on CASIA-Webface using ResNet50 as feature extractors. We trained models on multiple setups of FlatFace. For baselines, we also reproduced the experiments on conventional ArcFace and AdaFace. We did the experiments three times for all setups and report the average accuracies along with standard deviations in Table 2. As seen in the table, our methods outperform their baseline counterparts, with FlatFace-AACE-AdaFace achieving the highest average accuracy, improving by 0.29% over AdaFace.

Furthermore, We observe the generalization gaps, the difference between training accuracies and average testing accuracies, of the trained models. As presented in Fig. 2, FlatFace framework significantly reduce the generalization gaps compared to the standard FR methods.

Results on MS1MV2

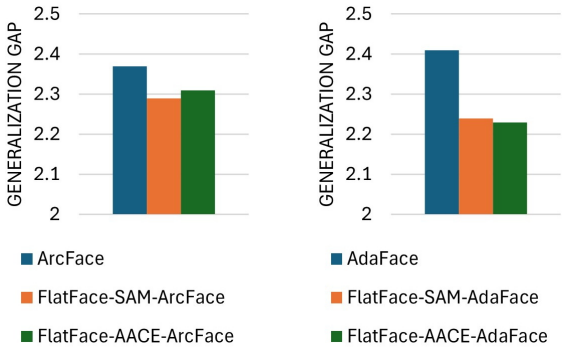
To test the performance of our methods on a large-scale dataset, we also employed the experiments on MS1MV2 dataset using ResNet100 as backbones for feature extraction. Table 3 presents the results of the experiments. Since the performances on MS1MV2 dataset are near saturated, the trained models show only slightly different results. However, FlatFace-SAM-AdaFace shows the highest average accuracy at $97.21 \pm 0.06\%$.

Similar to previous experiments, we also investigate the generalization gaps. As presented in Fig. 3, FlatFace variants still have lower generalization gaps compared to their conventional counterparts, which indicate better generalization.



(a) ArcFace-based experiments. (b) AdaFace-based experiments.

Figure 2: Generalization gaps comparison between normal margin-based FR and FlatFace trained on CASIA dataset.



(a) ArcFace-based experiments. (b) AdaFace-based experiments.

Figure 3: Generalization gaps comparison between normal margin-based FR and FlatFace trained on MS1MV2 dataset.

These empirical experiments confirm the decrease in generalization gaps by adopting FlatFace, which leads to improvement in generalization and overall performance.

Conclusion

In this work, we propose a new training framework named FlatFace that modifies the training pipeline of the face recognition methods by adopting SAM's weight perturbation techniques. This training framework consists of two main steps including the perturbation step that perturbs the model parameter toward the worst-case position, and the weight updating step that uses the loss gradient at the perturbed position to update model parameters. FlatFace guides the model toward flatter regions in the loss landscape, improving model generalization and accuracy in open-set face recognition tasks. The empirical experiments confirm the decrease in the generalization gap and the improvement in overall performance.

Acknowledgements

This work was partially supported by JSPS KAKENHI Grant Numbers 24K02957 and was carried out using the TSUBAME4.0 supercomputer at Institute of Science Tokyo.

Table 2: Accuracy comparison of models trained on CASIA dataset, tested with different benchmarks

Setup	LFW	CFP-FP	AgeDB	CPLFW	CALFW	Average
ArcFace	99.39 ± 0.05	96.99 ± 0.01	94.56 ± 0.31	89.46 ± 0.13	93.65 ± 0.19	94.81 ± 0.10
AdaFace	99.35 ± 0.02	96.84 ± 0.33	94.59 ± 0.15	90.24 ± 0.30	93.67 ± 0.03	94.94 ± 0.08
FlatFace-SAM-ArcFace	$\mathbf{99.49 \pm 0.01}$	97.24 ± 0.16	94.79 ± 0.15	89.78 ± 0.16	$\mathbf{93.94 \pm 0.05}$	95.05 ± 0.03
FlatFace-SAM-AdaFace	99.39 ± 0.07	96.90 ± 0.29	94.80 ± 0.50	90.44 ± 0.50	93.50 ± 0.57	95.01 ± 0.37
FlatFace-AACE-ArcFace	99.30 ± 0.07	$\mathbf{97.28 \pm 0.15}$	94.63 ± 0.10	89.84 ± 0.05	93.79 ± 0.06	94.97 ± 0.06
FlatFace-AACE-AdaFace	99.38 ± 0.02	97.24 ± 0.07	$\mathbf{94.98 \pm 0.30}$	$\mathbf{90.59 \pm 0.27}$	$\mathbf{93.94 \pm 0.13}$	$\mathbf{95.23 \pm 0.10}$

Table 3: Accuracy comparison of models trained on MS1MV2 dataset, tested with different benchmarks

Setup	LFW	CFP-FP	AgeDB	CPLFW	CALFW	Average
ArcFace	99.78 ± 0.04	$\mathbf{98.55 \pm 0.05}$	98.10 ± 0.07	92.98 ± 0.12	$\mathbf{96.19 \pm 0.01}$	97.12 ± 0.03
AdaFace	99.80 ± 0.02	98.34 ± 0.19	98.15 ± 0.20	93.03 ± 0.37	96.02 ± 0.10	97.07 ± 0.10
FlatFace-SAM-ArcFace	$\mathbf{99.81 \pm 0.01}$	98.40 ± 0.19	$\mathbf{98.24 \pm 0.08}$	92.79 ± 0.11	96.07 ± 0.09	97.06 ± 0.05
FlatFace-SAM-AdaFace	99.79 ± 0.02	98.48 ± 0.05	98.19 ± 0.03	$\mathbf{93.41 \pm 0.17}$	96.16 ± 0.08	$\mathbf{97.21 \pm 0.06}$
FlatFace-AACE-ArcFace	99.79 ± 0.01	98.48 ± 0.09	98.06 ± 0.18	93.18 ± 0.05	96.08 ± 0.16	97.12 ± 0.07
FlatFace-AACE-AdaFace	99.78 ± 0.02	98.44 ± 0.16	98.10 ± 0.02	93.20 ± 0.10	96.06 ± 0.12	97.12 ± 0.05

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[2] Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton, "Imagenet classification with deep convolutional neural networks," *Neural Information Processing Systems*, vol. 25, 01 2012.

[3] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 815–823, 2015.

[4] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," *Advances in neural information processing systems*, vol. 27, 2014.

[5] Jiankang Deng, Yuxiang Zhou, and Stefanos Zafeiriou, "Marginal loss for deep face recognition in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2006–2014, 2017.

[6] Rajeev Ranjan, Carlos D Castillo, and Rama Chellappa, "L2-constrained softmax loss for discriminative face verification," *arXiv preprint arXiv:1703.09507*, 2017.

[7] Changxing Ding and Dacheng Tao, "Robust face recognition via multimodal deep face representation," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 2049–2058, 2015.

[8] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 539–546 vol. 1, 2005.

[9] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network in *Similarity-based pattern recognition: third international workshop, SIMBAD 2015, Copenhagen, Denmark, October 12-14, 2015. Proceedings 3*. Springer, pp. 84–92, 2015.

[10] Weiyang Liu, Yandong Wen, Zhiding Yu, and Meng Yang, "Large-margin softmax loss for convolutional neural networks," *arXiv preprint arXiv:1612.02295*, 2016.

[11] Weiyang Liu, Yandong Wen, Zhiding Yu, Ming Li, Bhiksha Raj, and Le Song, "Sphereface: Deep hypersphere embedding for face recognition in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 212–220, 2017.

[12] 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*, pp. 5265–5274, 2018.

[13] 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*, pp. 4690–4699, 2019.

[14] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang, "Curricularface: adaptive curriculum learning loss for deep face recognition in *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5901–5910, 2020.

[15] Minchul Kim, Anil K Jain, and Xiaoming Liu, "Adaface: Quality adaptive margin for face recognition in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 18750–18759, 2022.

[16] Gintare Karolina Dziugaite and Daniel M Roy, "Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data," *arXiv preprint arXiv:1703.11008*, 2017.

[17] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," *arXiv preprint arXiv:1609.04836*, 2016.

[18] Sepp Hochreiter and Jürgen Schmidhuber, "Flat minima," *Neural computation*, vol. 9, no. 1, pp. 1–42, 1997.

[19] Pierre Foret, Ariel Kleiner, Hossein Mobahi, and Behnam Neyshabur, "Sharpness-aware minimization for efficiently improving generalization," *arXiv preprint arXiv:2010.01412*, 2020.

[20] Tanapat Ratchatorn and Masayuki Tanaka, "Adaptive adversarial cross-entropy loss for sharpness-aware minimization in *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 479–485, 2024.

[21] Herbert Robbins and Sutton Monro, "A stochastic approximation method," *The Annals of Mathematical Statistics*, vol. 22, no. 3, pp. 400–407, 1951.

[22] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li, "Learning face representation from scratch," *arXiv preprint arXiv:1411.7923*, 2014.

[23] Gary B. Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled Faces in the Wild: A Database forStudying Face Recognition in Unconstrained Environments in *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, Oct. 2008.

[24] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M. Patel, Rama Chellappa, and David W. Jacobs, "Frontal to profile face verification in the wild in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–9, 2016.

[25] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou, "Agedb: The first manually collected, in-the-wild age database in *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1997–2005, 2017.

[26] T. Zheng and W. Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," Tech. Rep. 18-01, Beijing University of Posts and Telecommunications, February 2018.

[27] Tianyue Zheng, Weihong Deng, and Jiani Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

Author Biography

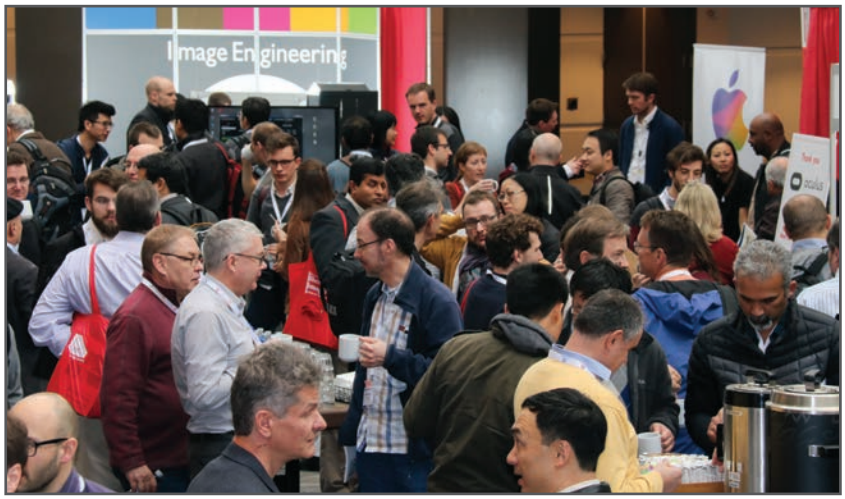
Tanapat Ratchatorn received a bachelor's degree in physics (2015) and a master's degree in electrical engineering (2020) from Chulalongkorn University, Thailand. From 2021 to 2023, he worked as a machine learning research engineer in the industry, focusing on computer vision applications. He was awarded the Thai government science and technology scholarship and joined the Institute of Science Tokyo (formerly Tokyo Institute of Technology), Japan, in 2023 as a Ph.D. student in systems and control engineering.

Masayuki Tanaka received his Ph.D. in control engineering from Tokyo Institute of Technology (2003). He was a Research Scientist (2004 to 2008) and Associate Professor (2008 to 2017, 2020 to 2023) at Tokyo Institute of Technology. He was a Visiting Scholar at Stanford University (2013 to 2014) and a Senior Researcher at the AIST (2017 to 2020). Since 2023, he has been a Professor at the Institute of Science Tokyo (formerly Tokyo Institute of Technology).

JOIN US AT THE NEXT EI!

electronic IMAGING

Imaging across applications . . . Where industry and academia meet!



• SHORT COURSES • EXHIBITS • DEMONSTRATION SESSION • PLENARY TALKS •
• INTERACTIVE PAPER SESSION • SPECIAL EVENTS • TECHNICAL SESSIONS •

www.electronicimaging.org

ADVANCING IMAGING SCIENCE 75 1947 IS&T 2022 imaging.org