TransFace: Calibrating Transformer Training for Face Recognition from a Data-Centric Perspective

Jun Dan [*1], Yang Liu [*2], Haoyu Xie[2], Jiankang Deng[3],
Haoran Xie[4], Xuansong Xie[2], Baigui Sun [†2]
[1]Zhejiang University [2]Alibaba DAMO Academy [3]Imperial College London [4]Tsinghua University
danjun@zju.edu.cn,j.deng16@imperial.ac.uk,xiehr20@mails.tsinghua.edu.cn
{ly261666, xiehaoyu.xhy,xingtong.xxs, baigui.sbg}@alibaba-inc.com

Abstract

Vision Transformers (ViTs) have demonstrated powerful representation ability in various visual tasks thanks to their intrinsic data-hungry nature. However, we unexpectedly find that ViTs perform vulnerably when applied to face recognition (FR) scenarios with extremely large datasets. We investigate the reasons for this phenomenon and discover that the existing data augmentation approach and hard sample mining strategy are incompatible with ViTs-based FR backbone due to the lack of tailored consideration on preserving face structural information and leveraging each local token information. To remedy these problems, this paper proposes a superior FR model called TransFace, which employs a patch-level data augmentation strategy named DPAP and a hard sample mining strategy named EHSM. Specially, DPAP randomly perturbs the amplitude information of dominant patches to expand sample diversity, which effectively alleviates the overfitting problem in ViTs. EHSM utilizes the information entropy in the local tokens to dynamically adjust the importance weight of easy and hard samples during training, leading to a more stable prediction. Experiments on several benchmarks demonstrate the superiority of our TransFace. Code and models are available at `https://github.com/DanJun6737/TransFace`.

1. Introduction

Over the past few years, Convolutional Neural Networks (CNNs) [23, 32] have achieved remarkable success in the computer vision community, thanks to the availability of large-scale datasets. Recently, the introduction of Vision Transformers (ViTs) [12] has caught the attention of the computer vision community due to their powerful representation abilities. Unlike CNN models, ViTs lack some convolution-like inductive biases, such as translation equivariance and locality, leading to challenges in the convergence of the ViTs backbone. To remedy this problem, pioneering works [61, 12, 4, 82] point out the rationale behind this derives from its data-hungry nature, indicating that a superior ViTs representation would be supported by large-scale training data. By taking advantage of their intrinsic data-hungry property, ViTs are commonly used to serve as an alternative backbone on several visual tasks [44, 12, 61, 58, 15, 17].

However, when considering an extremely data-adequate scenario to satisfy ViTs' data-hungry property, namely Face Recognition (FR), we unexpectedly discover that the performance of ViTs is almost equal to that of CNNs [80]. To explore why ViTs perform vulnerably in the FR realm, we investigate the training process of ViTs. From a data-centric perspective, we discover that the instance-level data augmentation approach and hard sample mining strategy are incompatible for ViTs-based FR backbone due to the lack of tailored consideration on preserving face structural information and leveraging each local token information (illustrated in Fig. 1). To handle these drawbacks, we make two following efforts: (i) Identify why and how to devise a patch-level data augmentation strategy on the ViTs-based FR backbone. (ii) Unveil why and how to mine representative token information.

Patch-level Data Augmentation Strategy. Due to the lack of inductive biases, ViT-based models are hard to train and prone to overfitting [61, 12, 4]. To alleviate the overfitting phenomenon, existing works [61, 82, 70] attempt several data augmentation strategies, such as Random Erasing [81], Mixup [76], CutMix [75], RandAugment [7] and their variants [4, 19, 61, 38], to construct diverse training samples. However, these instance-level data augmentation strategies are not suitable for the FR task, because they will inevitably

[*] Equal Contribution, [†] Corresponding Author .

destroy some key structural information of the face identity (as shown in the top of Fig. 1), which may lead the ViTs to optimize in the incorrect direction. Furthermore, the recent study [82] observes that ViTs are actually prone to overfitting to certain local patches during training, resulting in severely impaired generalization performance of the model. For example, in the FR task, the prediction of ViT may be dominated by a few face patches (*e.g.*, eyes and forehead). Therefore, once these key patches are disturbed (*e.g.*, a superstar wearing sunglasses or a hat), the model tends to make spurious decisions. These problems seriously affect the large-scale deployment of ViT-based FR models in real scenarios.

To solve the aforementioned problem, motivated by the structural information preserving property of the Fourier phase spectrum [50, 52, 73, 49], we introduce a patch-level data augmentation strategy named Dominant Patch Amplitude Perturbation (**DPAP**). Without destroying the fidelity and structural information of the face, DPAP can efficiently expand sample diversity. Concretely, DPAP uses a Squeeze-and-Excitation (SE) module [24] to screen out the top-K patches (**dominant patches**), then randomly mixes their amplitude information, and combines it with the original phase information to generate diverse samples.

Different from previous data augmentation strategies, the proposed DPAP cleverly utilizes the prior knowledge (*i.e.*, the position of dominant patches) provided by the model to augment data, which can more precisely alleviate the overfitting problem in ViTs. Furthermore, as diverse patches are continuously generated, DPAP also indirectly encourages ViTs to utilize other face patches, especially some patches that are easily ignored by the deep network (*e.g.*, ears, mouth, and nose), to make more confident decisions.

Hard Sample Mining Strategy. As demonstrated in Refs. [36, 26, 3], hard sample mining technology plays an important role in boosting the model's final performance via continuously assimilating knowledge from effective/hard samples. Most previous works are specially designed for CNNs, they usually adopt several instance-level indicators of the sample, such as prediction probability [36, 27], prediction loss [14, 56], and latent features [53], to mine hard samples (as shown in the bottom of Fig. 1). However, the recent study [82] has shown that the prediction of ViT is mainly determined by only a few patch tokens, which means that the global token of ViTs may be dominated by a few local tokens. Therefore, directly using such biased indicators to mine hard sample is suboptimal for ViTs, especially when some dominant local tokens are ignored.

To better mine hard samples, inspired by the information theory [51, 55, 5], we propose a novel hard sample mining strategy named Entropy-guided Hard Sample Mining (**EHSM**). EHSM treats the ViT as an information processing system, which dynamically adjusts the importance

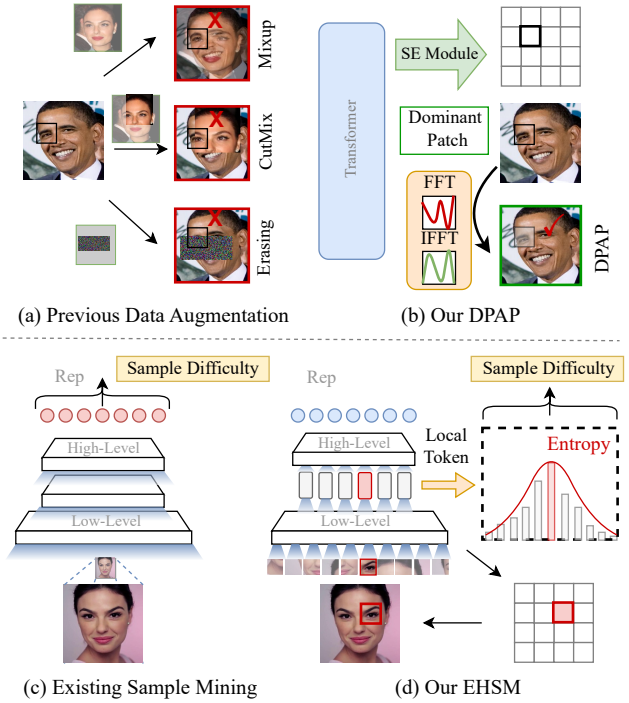


Figure 1. (Top) Previous data augmentation approaches may destroy the fidelity and structural information of face identity when augmenting samples. Our DPAP strategy not only constructs diverse samples but also effectively preserves the key information of the face. (Bottom) Existing hard sample mining methods usually adopt several instance-level indicators to measure sample difficulty, which is suboptimal for ViTs. Our EHSM strategy leverages information entropy from all local tokens to mine hard samples.

weight of easy and hard samples in terms of the total amount of information contained in the local tokens. It is worth mentioning that EHSM has the potential to encourage the ViT to fully mine fine-grained information contained in each face patch, especially some less-attended face cues (*e.g.*, lip and jaw), which greatly enhances the feature representation power of each local token (as verified by our experiment in Fig. 2 and Fig. 5). In this way, even if some important image patches are destroyed, the model can also make full use of the remaining face cues to generalize the global token, leading to a more stable prediction.

The following are the main contributions of this paper:

(1) A patch-level data augmentation strategy named DPAP is introduced to effectively alleviate the overfitting problem in ViTs.

(2) A novel hard sample mining strategy named EHSM is proposed to enhance the stability of FR model prediction.

(3) Experimental results on various popular face benchmarks have shown the superiority of our method, *e.g.*, we achieve 97.61% accuracy on "TAR@FAR=1E-4" of the

IJB-C benchmark using the Glint360K training set.

2. Related Works

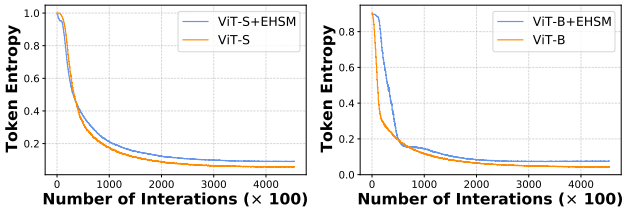
Vision Transformer (ViT). Recently, ViT has demonstrated its powerful feature representation ability in various vision tasks, including image recognition[12, 61, 62, 44], semantic segmentation[58, 4], and object detection/localization [15, 17]. Unlike CNN, ViT is mainly based on the self-attention mechanism [63], which can effectively capture the relationships between different features. It has been demonstrated that ViT does not generalize well when trained on insufficient amounts of training samples [12]. DeiT [61] introduces a novel knowledge distillation procedure to strengthen the ViT's generalization ability. To better extract both global and local visual information, TNT [22] employs an inner transformer block to model the relationship among sub-patches. ATS [16] is proposed to design computationally efficient ViT models by adaptively sampling significant tokens.

Moreover, several recently proposed studies aim to balance ViT's computation and accuracy. UniFormer [34] superbly unifies 3D convolution and spatiotemporal self-attention, significantly alleviating the computational burden in capturing token relation. Dilateformer [29] uses a sliding window to select representative patches, brilliantly mitigating the computational cost of self-attention. Ref. [30] proposes applying a masking mechanism into the attention map, tremendously reducing computational load between tokens.

Face Recognition (FR). CNNs have made significant progress in face-related tasks [43, 40, 42, 11, 10, 41]. Among them, extracting deep face embedding attracts many reaserchers' attention. There are two main ways to train CNNs for FR. One kind of way is metric learning-based methods, which aim to learn discriminative face representation, such as Triplet loss [53], Tuplet loss [57] and Center loss [68]. Another way is margin-based softmax methods, which focus on incorporating margin penalty into softmax classification loss framework, such as ArcFace [11, 9], CosFace [65], AM-softmax [64] and SphereFace [39]. To further improve the efficiency of margin-based softmax loss on the large-scale dataset, some studies' emphasis has shifted to adpative parameters [78, 77, 37, 31, 47], inter-class regularization [79, 13], sample mining [27, 67], learning acceleration [1, 35], knowledge distillation [26], etc.

The recent proposed Face transformer [80] first demonstrates the feasibility of using ViT in FR. However, there is still a lack of exploration on how to train a superior ViT-based FR model on the extremely large-scale dataset.

Data-Centirc ViTs. To improve the performance of ViTs, previous works [72, 22, 69, 74] mainly try to modify the structure of the models, which is very dependent on experience. Recently, several works have been proposed to enhance the generalization ability of ViTs from the view of training data. Gong *et al*. [18] proposes two patch-related loss functions to alleviate the over-smoothing problem in ViTs, which can promote the stable training of deeper ViTs without any structural modification. DeiT [61] shows that strong data augmentation strategies [76, 81, 7, 6] can help ViTs absorb data more efficiently. TransMix [4] introduces a data augmentation strategy that leverages the transformer's attention maps to guide the mixture of labels. TokenMix [38] proposes to mix images at the token level, which facilitates ViTs to infer sample classes more accurately.



(a) (b)

Figure 2. We visualize the trend of the average information entropy contained in the local token during training. With the help of EHSM, the face information contained in each patch is more fully mined and utilized. More results can be found in Fig. 5.

3. Methodology

3.1. Preliminaries

A classical ViT [12] first splits an input image into a sequence of fixed-size patches. Each small patch is mapped to a feature vector (or called a token) by a linear layer. Then, an additional learnable class token is concatenated to the tokens and position embeddings are added to each token to retain positional information. After that, the mixed embeddings are sent into a transformer encoder for feature encoding. More specially, the standard transformer encoder consists of multiheaded self-attention (MSA) and MLP blocks. LayerNorm (LN) and residual connections are applied before and after each block, respectively. Finally, the class token of the transformer encoder output is selected as the final representation and fed into a classification head for prediction.

Different from the original ViT architecture, we follow insightface[1] and do not employ the learnable class token in our model. The architecture of our TransFace model is depicted in Fig. 3. It is made up of three components: a transformer encoder $\mathcal{F}$, a "Squeeze-and-Excitation" (SE) module $\mathcal{S}$ [24] and a classification head $\mathcal{C}$. We will detail the architecture of our model in section 3.2.

[1]https://github.com/deepinsight/insightface/tree/master/recognition

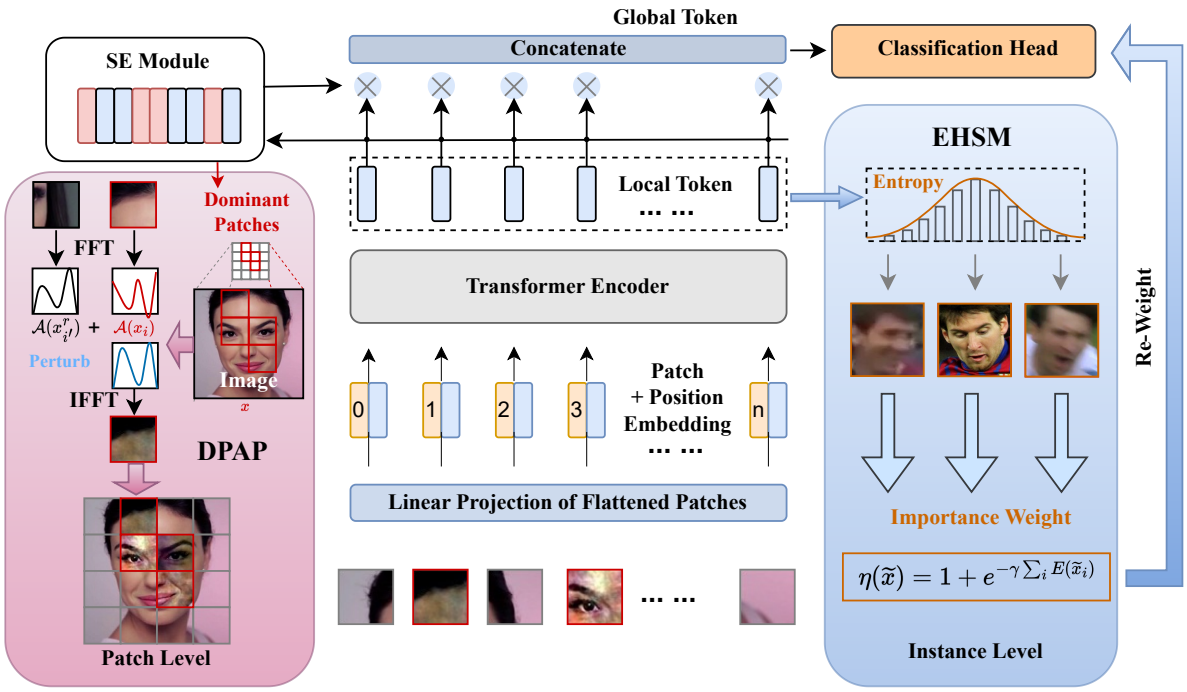


Figure 3. Global overview of our proposed TransFace model. To alleviate the overfitting problem in ViTs, the DPAP strategy employs the SE module to screen out the top-K dominant patches, and then randomly perturbs their amplitude information to expand sample diversity. Furthermore, to better mine hard samples and enhance the feature presentation power of local tokens, the EHSM strategy utilizes an entropy-aware weight mechanism to re-weight the classification loss. n is the total number of patches, and $\bigotimes$ denotes the multiplication operation between the local token and the scaling factor generated by the SE module. The image patches with red boxes represent dominant patches.

3.2. Patch-Level Data Augmentation Strategy

Existing works mainly apply a series of instance-level data augmentation strategies to alleviate the overfitting phenomenon in ViTs. However, these strategies will inevitably destroy some key structural information of face identity (as shown in Fig. 1), which may seriously affect the learning of discriminative face tokens. Furthermore, the recent study [82] observes that ViTs are actually prone to overfitting to a few patches, which greatly limits the deployment and scalability of ViT-based FR models in application scenarios.

To address the aforementioned issues, a novel patch-level data augmentation strategy named Dominant Patch Amplitude Perturbation (**DPAP**) is proposed for the ViTs-based FR backbone. The main steps of the strategy are: (1) First, we insert an SE module $\mathcal{S}$ at the output of the transformer encoder $\mathcal{F}$, and utilize the scaling factors generated by $\mathcal{S}$ to find out the top-K patches (*i.e.*, **dominant patches**) of the original image x that contribute the most to the final prediction. (2) Second, we employ a linear mixing mechanism to randomly perturb the amplitude information of these dominant patches. (3) Finally, we feed the reconstructed image $\widetilde{x}$ into our TransFace model for supervised training.

Mathematically, given an image x, we denote a sequence of decomposed image patches as $x = (x_1, x_2, \cdots, x_n)$, where each x_i represents an image patch and n is the total number of patches. And the output of the transformer encoder $\mathcal{F}$ is denoted as $(f_1, f_2, \cdots, f_n)$, where $f_1, \cdots, f_n$ represent the local tokens. Then, all the local tokens extracted by $\mathcal{F}$ will pass through the SE module $\mathcal{S}$ and be rescaled as $(\kappa_1 \cdot f_1, \kappa_2 \cdot f_2, \cdots, \kappa_n \cdot f_n)$, where $\kappa_1, \cdots, \kappa_n$ denote the scaling factors generated by $\mathcal{S}$. Actually, these scaling factors $\kappa_1, \cdots, \kappa_n$ of local tokens indirectly reflect the importance of local tokens in prediction. We further normalize these scaling factors using the softmax function:

$$d_i = \frac{e^{\kappa_i}}{\sum_{i'} e^{\kappa_{i'}}}. \tag{1}$$

According to the top-K largest normalized scaling factors, we can screen out the top-K dominant patches that the model "cares about" the most. To relieve the ViTs from overfitting to these dominant patches, a natural idea is to let the model "see" more diverse patches. Motivated by the structural information preserving property of the Fourier phase spectrum [50, 52, 73, 49, 71], we propose to perturb the amplitude spectrum information of their dominant

patches using a linear mix mechanism and keep their phase spectrum information unchanged.

For a single channel image patch x_i, its Fourier Transform $\mathcal{T}(x_i)$ can be expressed as:

$$\mathcal{T}(x_i)(u,v) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} x_i(h,w)e^{-j2\pi(\frac{h}{H}u+\frac{w}{W}v)}, \quad (2)$$

where $j^2=-1$, H and W represent the height and width of x_i, respectively. $\mathcal{T}^{-1}(x_i)$ represents the corresponding inverse Fourier Transform that maps the amplitude and phase information back to the original image space. The Fourier transform and its inverse transform can be efficiently implemented by the FFT and IFFT algorithms [2], respectively.

$\mathcal{A}(x_i)$ and $\mathcal{P}(x_i)$ represent the amplitude spectrum and phase spectrum of the image patch x_i, respectively:

$$\begin{aligned}\mathcal{A}(x_i)(u,v) &= \left[\mathcal{R}^2(x_i)(u,v)+\mathcal{I}^2(x_i)(u,v)\right]^{1/2},\\ \mathcal{P}(x_i)(u,v) &= \arctan\left[\frac{\mathcal{I}(x_i)(u,v)}{\mathcal{R}(x_i)(u,v)}\right],\end{aligned} \quad (3)$$

where $\mathcal{R}(x_i)$ and $\mathcal{I}(x_i)$ denote the real and imaginary part of $\mathcal{T}(x_i)$, respectively. For RGB face image patches, we need to calculate the Fourier Transform of each channel independently to obtain the final amplitude spectrum and phase spectrum.

To effectively construct diverse images without destroying the fidelity and structural information of face identity, we employ a Mixup-like mechanism to linearly mix the amplitude spectrum of dominant patch x_i and random patch $x_{i'}^r$:

$$\widetilde{\mathcal{A}}(x_i) = \lambda\mathcal{A}(x_i) + (1-\lambda)\mathcal{A}(x_{i'}^r), \quad (4)$$

where $\lambda \sim U(0,\alpha)$, $U(0,\alpha)$ is a uniform distribution on $[0,\alpha]$, $x_{i'}^r$ is a random patches of a random training sample x^r, and α is a hyper-parameter used to control the intensity of amplitude information mixing. We then combine the mixed amplitude spectrum with the original phase spectrum to reconstruct a new Fourier representation:

$$\mathcal{T}(\widetilde{x}_i)(u,v) = \widetilde{\mathcal{A}}(x_i)(u,v) * e^{-j*\mathcal{P}(x_i)(u,v)}, \quad (5)$$

which will be mapped to the original image space by the inverse Fourier transform to generate a new patch, *i.e.*, $\widetilde{x}_i = \mathcal{T}^{-1}\left[\mathcal{T}(\widetilde{x}_i)(u,v)\right]$. We then feed the augmented image $\widetilde{x}$ into the model for supervised training.

For the augmented image $\widetilde{x}$, we denote the local tokens extracted by $\mathcal{F}$ as $(\widetilde{f}_1, \widetilde{f}_2, \cdots, \widetilde{f}_n)$, which will be further fed into the SE module $\mathcal{S}$ and re-scaled as $(\widetilde{\kappa}_1 \cdot \widetilde{f}_1, \widetilde{\kappa}_2 \cdot \widetilde{f}_2, \cdots, \widetilde{\kappa}_n \cdot \widetilde{f}_n)$. Then, all the re-scaled local tokens are concatenated into a global token $\widetilde{g} = \left[\widetilde{\kappa}_1 \cdot \widetilde{f}_1; \widetilde{\kappa}_2 \cdot \widetilde{f}_2; \cdots; \widetilde{\kappa}_n \cdot \widetilde{f}_n\right]$, which will be used for the subsequent FR task via the classification head $\mathcal{C}$. In our model, we adopt the most widely used ArcFace Loss [11] as the basic classification loss:

$$\mathcal{L}_{arc}(\widetilde{g}, y) = -\log\frac{e^{s(\cos(\theta_y+m))}}{e^{s(\cos(\theta_y+m))}+\sum_{l=1,l\neq y}^{c} e^{s\cos\theta_l}}, \quad (6)$$

where y denotes the class label of image x, s is a scaling hyperparameter, θ_l is the angle between the l-th column weight and feature, $m>0$ denotes an additive angular margin and c is the class number.

As the dominant patches are constantly changed, the DPAP strategy can indirectly promote the FR model to utilize other face patches, especially some patches that are easily ignored by the deep network (*e.g.*, ears, mouth, and nose), to assist the final prediction. More importantly, the DPAP strategy cleverly utilizes the prior knowledge provided by the model (*i.e.*, the position of the dominant patches) to perform data augmentation, which can more effectively mitigate the overfitting problem and enhance the generalization ability of ViTs.

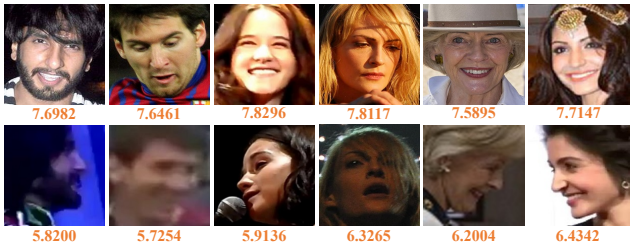


Figure 4. Example images and corresponding information entropy. Samples labeled with the same ID are displayed in each column. (First row) Easy samples usually contain richer information (*i.e.*, larger information entropy). (Second row) Hard samples usually contain less information (*i.e.*, lower information entropy).

3.3. Entropy-based Hard Sample Mining Strategy

As shown in Refs. [36, 26, 3], hard sample mining technology is often adopted to further boost the model's final performance. Previous works on hard sample mining, such as Focal loss [36], MV-Softmax [67], OHEM [56], AT_k loss [14], etc., are specially designed for CNNs, they aim to encourage models to explicitly emphasize the impact of hard samples [27, 53]. These methods usually utilize several instance-level indicators of the sample, such as prediction probability [36, 27], prediction loss [14, 56], and latent features [53], to directly or indirectly measure sample difficulty. However, the recent study [82] has demonstrated that the prediction of ViT is mainly determined by only a few patch tokens, which means that the global token of ViT may be dominated by several local tokens. Therefore, for ViTs, it is suboptimal to use such biased indicators to mine hard samples, especially when some dominant local tokens are ignored.

To mine hard samples more precisely, motivated by the information theory [51, 55, 5], we propose to measure sample difficulty in terms of the total amount of information contained in the local tokens. As demonstrated in Fig. 4, high-quality face images (easy samples) usually contain richer information (*i.e.*, higher information entropy) and thus are easier to be learned by the model. Low-quality face images (hard samples), such as blur face images and low-contrast face images, usually contain relatively less useful information (*i.e.*, lower information entropy), so they are more difficult to be learned.

When we regard a deep neural network $\mathcal{M}$ as an information processing system, we can abstract its topology into a graph $\mathcal{G} = (\mathcal{Z}, \mathcal{Q})$. A series of neurons form the vertex set $\mathcal{Z}$, and the connections between neurons form the edge set $\mathcal{Q}$. For any $z \in \mathcal{Z}$ and $q \in \mathcal{Q}$, $e(z)$ and $e(q)$ denote the values of each vertex z and each edge q, respectively. Thus, the continuous state space of the deep network $\mathcal{M}$ can be defined by the set $\Omega = \{e(z), e(q) : \forall z \in \mathcal{Z}, q \in \mathcal{Q}\}$. In this way, the total information contained in $\mathcal{M}$ can be measured by the entropy $E(\Omega)$ of set Ω. The sets $E(\Omega_z) = \{e(z) : z \in \mathcal{Z}\}$ and $E(\Omega_q) = \{e(q) : q \in \mathcal{Q}\}$ represent the total information contained in the latent features and in the network parameters, respectively. Specially, $E(\Omega_z)$ measures the feature representation power of network $\mathcal{M}$ and $E(\Omega_q)$ measures the network complexity. In our work, we focus on the entropy of latent features $E(\Omega_z)$ rather than the entropy of network parameters $E(\Omega_q)$.

However, in the deep network $\mathcal{M}$, the latent features of image always follow a complex and unknown distribution [20, 8], it is difficult to directly calculate the information entropy of latent features $E(\Omega_z)$. Fortunately, the Maximum Entropy Principle [5, 28, 33, 60] has proved that the entropy of a distribution is upper bounded by a Gaussian distribution with the same mean and variance, as shown in Theorem 1.

Theorem 1 *For any continuous distribution $\mathbb{D}(a)$ of mean μ and variance σ^2, its differential entropy is maximized when $\mathbb{D}(a)$ is a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$.*

Therefore, we can estimate the upper bound of the entropy instead. Suppose a is sampled from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$, the differential entropy of a can be defined as follows:

$$E(a) = \frac{1}{2}(\log(2\pi\sigma^2) + 1) = \frac{1}{2}(1 + \log(2\pi) + \log(\sigma^2)). \tag{7}$$

As can be seen, the entropy of the Gaussian distribution only depends on the variance. In this way, we can effectively approximate the entropy of the latent features $E(\Omega_z)$ by simply computing the variance of the latent features.

Different from previous works that adopt biased indicators to mine hard sample, we propose a novel hard sample mining strategy named Entropy-guided Hard Sample Mining (**EHSM**) to better achieve this goal. EHSM comprehensively considers the local and global information of tokens in measuring sample difficulty. Specifically, for an augmented sample $\widetilde{x} = (\widetilde{x}_1, \widetilde{x}_2, \cdots, \widetilde{x}_n)$, EHSM first estimates the local information entropy $E(\widetilde{x}_i) = E(\widetilde{\kappa}_i \cdot \widetilde{f}_i)$ of each local token $\widetilde{\kappa}_i \cdot \widetilde{f}_i$ using Equation (7). Then, all the local information entropy is aggregated as the global information entropy $\sum_i E(\widetilde{x}_i)$ of sample $\widetilde{x}$. Finally, EHSM utilizes an entropy-aware weight mechanism $\eta(\widetilde{x}) = 1 + e^{-\gamma \sum_i E(\widetilde{x}_i)}$ to adaptively assign importance weight to each sample, where γ is a temperature coefficient. The re-weighted classification loss can be formulated as:

$$\mathcal{L}_{cls} = \eta(\widetilde{x}) \times \mathcal{L}_{arc}(\widetilde{g}, y). \tag{8}$$

It is worth mentioning that EHSM explicitly encourages the model to focus on hard samples with less information. In order to minimize the objective $\mathcal{L}_{cls}$, the model has to optimize both weight η and basic classification loss $\mathcal{L}_{arc}$ during training, which will bring two benefits: (1) Minimizing $\mathcal{L}_{arc}$ can encourage the model to learn better face features from diverse training samples. (2) Minimizing the weight $\eta(\widetilde{x})$ (*i.e.*, maximizing the total information $\sum_i E(\widetilde{x}_i)$) will facilitate the model to fully mine feature information contained in each face patch, especially some less-attended face cues (*e.g.*, nose, lip, and jaw), which significantly enhances the feature representation power of each local token. In this way, even if some important face patches are destroyed, the model can also make full use of the remaining face cues to generalize the global token, leading to a more stable prediction.

Training Data	Method	GFLOPs	LFW	CFP-FP	AgeDB-30
Glint360K	R50, ArcFace	6.3	99.78	98.77	98.28
	R100, ArcFace	12.1	99.81	99.04	98.31
	R200, ArcFace	23.4	99.82	99.14	98.49
	ViT-S	5.7	99.80	98.85	98.24
	ViT-B	11.4	99.82	99.02	98.33
	ViT-L	25.3	99.82	99.10	98.47
	TransFace-S	5.8	**99.85**	98.91	98.50
	TransFace-B	11.5	**99.85**	99.17	98.53
	TransFace-L	25.4	**99.85**	**99.32**	**98.62**

Table 1. Verification accuracy (%) on LFW, CFP-FP and AgeDB-30 benchmarks.

4. Experiments

4.1. Implementation Details

Datasets. We separately adopt MS1MV2 [11] dataset (5.8M images, 85k identities) and the recently proposed larger-scale Glint360K [1] dataset (17M images, 360K identities) to train our model. For evaluation, we employ LFW[25], AgeDB-30 [48], CFP-FP [54] and IJB-C [46] as the benchmarks to test the recognition performance of our model.

Training Data	Method	GFLOPs	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)	IJB-C(1E-3)	IJB-C(1E-2)	IJB-C(1E-1)
MS1MV2	R100, Softmax [47]	12.1*	64.07	83.68	92.40	-	-	-
	R100, SV-AM-Softmax [66, 47]	12.1∗	63.65	80.30	88.34	-	-	-
	R100, SphereFace [39, 47]	12.1∗	68.86	83.33	91.77	-	-	-
	R100, CosFace [65, 47]	12.1∗	87.96	92.68	95.56	-	-	-
	R100, ArcFace [11]	12.1∗	-	-	95.60	-	-	-
	R100, MV-Arc-Softmax [67, 27]	12.1*	-	-	95.20	-	-	-
	R100, CircleLoss [59]	12.1∗	-	89.60	93.95	96.29	-	-
	R100, CurricularFace [27]	12.1∗	-	-	96.10	-	-	-
	R100, MagFace [47]	12.1∗	**89.26**	93.67	95.81	-	-	-
	ViT-S	5.7	86.14	93.40	95.89	97.24	98.21	98.80
	ViT-B	11.4	86.66	94.08	96.15	97.38	98.24	98.89
	ViT-L	25.3	86.77	94.11	96.24	97.42	98.26	98.94
	TransFace-S	5.8	86.75	93.87	96.45	97.51	98.34	98.99
	TransFace-B	11.5	86.73	94.15	96.55	97.73	**98.47**	**99.11**
	TransFace-L	25.4	86.90	**94.55**	**96.59**	**97.80**	98.45	99.04
Glint360K	R50, ArcFace	6.3	88.40	95.29	96.81	97.79	98.30	99.04
	R100, ArcFace	12.1	88.38	95.38	96.89	97.86	98.33	99.07
	R200, ArcFace	23.4	89.45	95.71	97.20	97.98	98.38	99.09
	ViT-S	5.7	88.52	95.24	96.70	97.71	98.29	99.01
	ViT-B	11.4	88.58	95.41	96.88	97.80	98.35	99.09
	ViT-L	25.3	89.69	95.78	97.13	97.91	98.43	99.09
	TransFace-S	5.8	**89.93**	96.06	97.33	98.00	98.49	99.11
	TransFace-B	11.5	88.64	96.18	97.45	98.17	**98.66**	**99.23**
	TransFace-L	25.4	89.71	**96.29**	**97.61**	**98.26**	98.64	99.19

Table 2. Verification accuracy (%) on IJB-C benchmark. ∗ denotes R100 GFLOPs under 112 × 112 resolution.

Training Settings. Our experiments are implemented using Pytorch on 8 NVIDIA Tesla V100 GPUs. We follow [11] to employ ArcFace (s = 64 and m = 0.5) as the basic classification loss and crop all the input images to 112×112 by RetinaFace [10, 21]. To optimize the ViT-based FR models, we apply AdamW [45] optimizer with a weight decay of 0.1 for better convergence. The base learning rate for MS1MV2 is set to 1e-3 and 1e-4 for Glint360K. The detailed architecture of ViT models can be found in insightface[2]. The SE module consists of two fully connected layers, each with 144 neurons, followed by ReLu and Sigmoid activation functions, respectively. Note that all the models are learned from scratch without pre-training. In practice, we adopt variance instead of entropy to measure the amount of information contained in each local token to optimize the models more stably. For the hyperparameter α in DPAP, we choose $\alpha = 1$ for all experiments.

4.2. Results on Mainstream Benchmarks

Results on LFW, CFP-FP, and AgeDB-30. We train our TransFace on Glint360K, and compare it with other methods on various benchmarks, as reported in Table 1. We can find that the performance of the ViT baseline is comparable to that of the ResNet-based model. Notably, the accuracy of our TransFace model on these benchmarks is already near saturation. Specially, TransFace-L is higher than ViT-L by +0.03%, +0.22%, and +0.15% on three datasets, respectively.

Results on IJB-C. We train our TransFace on MS1MV2 and Glint360K respectively, and compare with SOTA competitors on the IJB-C benchmark, as reported in Table 2. We can observe that our three TransFace models trained with the MS1MV2 dataset greatly beat other ResNet-based models on "TAR@FAR=1E-4". For example, compared to CurricularFace, TransFace-B achieves +0.45% improvement on "TAR@FAR=1E-4". Furthermore, TransFace-S outperforms ViT-S by +0.56% on "TAR@FAR=1E-4".

On the Glint360K training set, our models significantly outperform other competitors. Specially, TransFace-L obtains the overall best results and greatly surpasses ViT-L by +0.48% and +0.51% on "TAR@FAR=1E-4" and "TAR@FAR=1E-5", respectively. These improved results demonstrate the superiority of our TransFace.

[2]https://github.com/deepinsight/insightface/tree/master/recognition

Method	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)
ViT-S	86.14	93.40	95.89
ViT-S + SE	86.26	93.76	96.12
ViT-S + DPAP	86.60	93.82	96.30
TransFace-S	**86.75**	**93.87**	**96.45**

Table 3. Ablation study of our model. Training Data: MS1MV2.

4.3. Analysis and Ablation Study

1) Contribution of Each Component: To investigate the contribution of each component in our model, we employ MS1MV2 as the training set and compare TransFace-S, ViT-S (baseline), and two variants of TransFace-S on the IJB-C benchmark. The variants of TransFace-S are as follows: (1) ViT-S + SE, the variant only introduces the SE module in the ViT-S model. (2) ViT-S + DPAP, based on ViT-S, the variant adds the DPAP strategy.

Method	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)
ViT-S	86.14	93.40	95.89
ViT-S + Random Erasing	83.68	93.27	96.06
ViT-S + RandAugment	86.24	93.74	96.14
ViT-S + PatchErasing	86.26	93.65	96.11
ViT-S + Mixup	85.51	93.41	96.08
ViT-S + CutMix	85.30	93.53	96.12
ViT-S + DPAP	**86.60**	**93.82**	**96.30**

Table 4. Comparison with previous data augmentation strategies. Training Data: MS1MV2.

Method	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)
ViT-S + AT_k	86.24	93.74	96.03
ViT-S + MV-Softmax	86.26	93.73	96.08
ViT-S + Focal loss	86.14	93.71	96.11
ViT-S + EHSM (global)	86.41	93.76	96.13
ViT-S + EHSM	**86.46**	**93.85**	**96.22**

Table 5. Comparison with previous hard sample mining strategies. Training Data: MS1MV2.

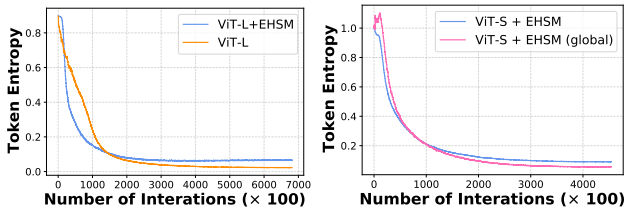
Parameter K	Method	IJB-C(1E-6)	IJB-C(1E-5)	IJB-C(1E-4)
1	TransFace-S	85.28	93.72	96.38
3	TransFace-S	86.30	93.72	96.42
5	TransFace-S	85.50	93.78	96.43
7	TransFace-S	**86.75**	**93.87**	**96.45**
10	TransFace-S	86.30	93.72	96.42
20	TransFace-S	84.90	93.65	96.35

Table 6. Parameter sensitivity analysis of our model on IJB-C. Training Data: MS1MV2.

The results gathered in Table 3 reflect the following observations: (1) Compared with ViT-S, the accuracy of ViT-S + SE is slightly improved due to the addition of the SE module. (2) ViT-S + DPAP outperforms ViT-S + SE, which indicates that perturbing the amplitude spectrum of dominant patches can effectively alleviate the overfitting problem in ViTs. (3) TransFace-S works better than ViT-S + DPAP, which shows the effectiveness of our EHSM strategy.

2) Comparison with Previous Data Augmentation Strategies: To further demonstrate the superiority of our DPAP strategy, we compare it with existing data augmentation strategies, including Random Erasing [81], Mixup [76], CutMix [75], RandAugment [7] and the recently proposed PatchErasing [82]. We employ MS1MV2 to train models and evaluate their performance on IJB-C. As reported in Table 4, compared with other strategies, our DPAP strategy can bring greater performance gain to the ViT, which benefits from the utilization of prior knowledge (*i.e.*, the position of dominant patches) and the preservation of structural information of face identity.

3) Does EHSM Enhance the Feature Representation Power of Each Token? We investigate the trend of the average information entropy contained in the local token of ViT (baseline) and variant ViT + EHSM during training on the MS1MV2 dataset, as shown in Figs. 2 and 5a. We can find that with the addition of our EHSM strategy, the token-level information becomes richer, which demonstrates the superiority of the EHSM strategy in improving the feature representation power of each local token.



(a) (b)

Figure 5. The trend of the average information entropy contained in the local token during training. With the help of EHSM, the face information contained in each patch is more fully mined and utilized.

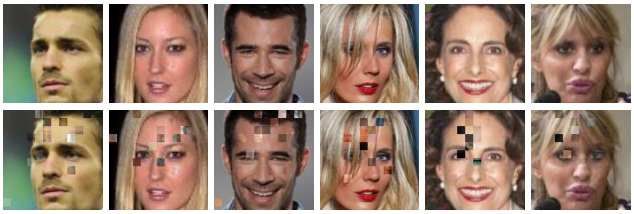


Figure 6. (First row) Original training samples. (Second row) Training samples augmented by DPAP strategy.

4) Effectiveness of EHSM: To demonstrate the superiority of EHSM in mining hard samples, we compare it with previous strategies, including AT_k loss [14], MV-Softmax [67] and Focal loss [36], as reported in Table 5. We can observe that the proposed EHSM strategy significantly outperforms previous hard sample mining strategies, which indicates that our EHSM strategy can better measure sample difficulty and boost the model's performance.

Moreover, in order to validate the advantages of the EHSM strategy in mining hard samples by comprehensively leveraging both local and global information, we conduct a further comparison between ViT-S + EHSM and its variant ViT-S + EHSM (global) that directly utilizes the entropy of the global token to measure sample difficulty. The results gathered in Table 5 demonstrate that ViT-S + EHSM greatly outperforms variant ViT-S + EHSM (global), which indicates that fully combining the information entropy of all the local tokens can more comprehensively measure sample difficulty than only using the entropy of the global token. Furthermore, compared to variant ViT-S + EHSM (global), ViT-S + EHSM can more effectively enhance the feature representation power of each local token, as illustrated in Fig. 5b.

5) Visualization of DPAP: As shown in Fig. 6, we visualize the original training samples and the samples augmented by the DPAP strategy ($K = 15$) during training on MS1MV2. We can see that the dominant patches are mainly distributed near hair, forehead, and eyes, which conforms with our in-

tuitions. The proposed DAPA strategy effectively relieves the model from overfitting to these dominant patches by perturbing their Fourier amplitude information, which indirectly encourages ViTs to utilize the remaining face cues (*e.g*., nose, mouth, ears, and jaw) to assist the final prediction, significantly enhancing the model's generalization ability.

6) Parameter Sensitivity: To investigate the effect of parameter K (*i.e*., the number of selected dominant patches) in our model, we adopt the MS1MV2 dataset to train TransFace-S with different K and evaluate their performance on IJB-C. The results gathered in Table 6 indicate that our model is robust to K. As K increases, the overall accuracy first rises and then falls, which verifies that properly perturbing the amplitude information of dominant patches can effectively alleviate the overfitting problem.

5. Conclusion

In this paper, we develop a novel model called TransFace to rescue the vulnerable performance of ViTs in the FR task. Specially, we introduce a patch-level data augmentation strategy named DPAP and a hard sample mining strategy named EHSM. Among them, DPAP adopts a linear mix mechanism to perturb the amplitude information of dominant patches to alleviate the overfitting problem in ViTs. EHSM fully utilizes the information entropy of multiple local tokens to measure sample difficulty, greatly enhancing the feature representation power of local tokens. Beyond the addition of the SE module, TransFace does not introduce any significant architectural change. Comprehensive experiments on popular facial benchmarks verify the superiority of TransFace. We hope our findings can shed some light on future research on ViTs-based FR as well as several relevant topics, *e.g*., personalized text-to-image generation model (AIGC) and 3D face reconstruction.

References

[1] Xiang An, Jiankang Deng, Jia Guo, Ziyong Feng, XuHan Zhu, Jing Yang, and Tongliang Liu. Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4042–4051, 2022.

[2] E Oran Brigham and RE Morrow. The fast fourier transform. *IEEE spectrum*, 4(12):63–70, 1967.

[3] Beidi Chen, Weiyang Liu, Zhiding Yu, Jan Kautz, Anshumali Shrivastava, Animesh Garg, and Animashree Anandkumar. Angular visual hardness. In *International Conference on Machine Learning*, pages 1637–1648. PMLR, 2020.

[4] Jie-Neng Chen, Shuyang Sun, Ju He, Philip HS Torr, Alan Yuille, and Song Bai. Transmix: Attend to mix for vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2022.

[5] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[6] Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation policies from data. *arXiv preprint arXiv:1805.09501*, 2018.

[7] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

[8] Jun Dan, Tao Jin, Hao Chi, Yixuan Shen, Jiawang Yu, and Jinhai Zhou. Homda: High-order moment-based domain alignment for unsupervised domain adaptation. *Knowledge-Based Systems*, 261:110205, 2023.

[9] Jiankang Deng, Jia Guo, Tongliang Liu, Mingming Gong, and Stefanos Zafeiriou. Sub-center arcface: Boosting face recognition by large-scale noisy web faces. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 741–757. Springer, 2020.

[10] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020.

[11] 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.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[13] Yueqi Duan, Jiwen Lu, and Jie Zhou. Uniformface: Learning deep equidistributed representation for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3415–3424, 2019.

[14] Yanbo Fan, Siwei Lyu, Yiming Ying, and Baogang Hu. Learning with average top-k loss. *Advances in neural information processing systems*, 30, 2017.

[15] Yuxin Fang, Bencheng Liao, Xinggang Wang, Jiemin Fang, Jiyang Qi, Rui Wu, Jianwei Niu, and Wenyu Liu. You only look at one sequence: Rethinking transformer in vision through object detection. *Advances in Neural Information Processing Systems*, 34:26183–26197, 2021.

[16] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Jürgen Gall. Adaptive token sampling for efficient vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XI*, pages 396–414. Springer, 2022.

[17] Wei Gao, Fang Wan, Xingjia Pan, Zhiliang Peng, Qi Tian, Zhenjun Han, Bolei Zhou, and Qixiang Ye. Ts-cam: Token

semantic coupled attention map for weakly supervised object localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2886–2895, 2021.

[18] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Improve vision transformers training by suppressing over-smoothing. *arXiv preprint arXiv:2104.12753*, 4(11), 2021.

[19] Chengyue Gong, Dilin Wang, Meng Li, Vikas Chandra, and Qiang Liu. Vision transformers with patch diversification. *arXiv preprint arXiv:2104.12753*, 2021.

[20] Mengran Gou, Octavia Camps, and Mario Sznaier. mom: Mean of moments feature for person re-identification. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 1294–1303, 2017.

[21] Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations*, 2022.

[22] Kai Han, An Xiao, Enhua Wu, Jianyuan Guo, Chunjing Xu, and Yunhe Wang. Transformer in transformer. *Advances in Neural Information Processing Systems*, 34:15908–15919, 2021.

[23] 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*, pages 770–778, 2016.

[24] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018.

[25] 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*, 2008.

[26] Yuge Huang, Pengcheng Shen, Ying Tai, Shaoxin Li, Xiaoming Liu, Jilin Li, Feiyue Huang, and Rongrong Ji. Improving face recognition from hard samples via distribution distillation loss. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 138–154. Springer, 2020.

[27] 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*, pages 5901–5910, 2020.

[28] Edwin T Jaynes. Information theory and statistical mechanics. *Physical review*, 106(4):620, 1957.

[29] Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Jinhua Ma, Yaowei Wang, and Wei-Shi Zheng. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE Transactions on Multimedia*, 2023.

[30] Kyungmin Kim, Bichen Wu, Xiaoliang Dai, Peizhao Zhang, Zhicheng Yan, Peter Vajda, and Seon Joo Kim. Rethinking the self-attention in vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3071–3075, 2021.

[31] 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*, pages 18750–18759, 2022.

[32] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, 2017.

[33] Solomon Kullback. *Information theory and statistics*. Courier Corporation, 1997.

[34] Kunchang Li, Yali Wang, Gao Peng, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatial-temporal representation learning. In *International Conference on Learning Representations*, 2021.

[35] Pengyu Li, Biao Wang, and Lei Zhang. Virtual fully-connected layer: Training a large-scale face recognition dataset with limited computational resources. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13315–13324, 2021.

[36] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[37] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z Li. Adaptiveface: Adaptive margin and sampling for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11947–11956, 2019.

[38] Jihao Liu, Boxiao Liu, Hang Zhou, Hongsheng Li, and Yu Liu. Tokenmix: Rethinking image mixing for data augmentation in vision transformers. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVI*, pages 455–471. Springer, 2022.

[39] 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*, pages 212–220, 2017.

[40] Yang Liu, Jiankang Deng, Fei Wang, Lei Shang, Xuansong Xie, and Baigui Sun. Damofd: Digging into backbone design on face detection. In *The Eleventh International Conference on Learning Representations*, 2022.

[41] Yang Liu and Xu Tang. Bfbox: Searching face-appropriate backbone and feature pyramid network for face detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13568–13577, 2020.

[42] Yang Liu, Xu Tang, Junyu Han, Jingtuo Liu, Dinger Rui, and Xiang Wu. Hambox: Delving into mining high-quality anchors on face detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13043–13051. IEEE, 2020.

[43] Yang Liu, Fei Wang, Jiankang Deng, Zhipeng Zhou, Baigui Sun, and Hao Li. Mogface: Towards a deeper appreciation on face detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4093–4102, 2022.

[44] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer:

Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[46] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. Iarpa janus benchmark-c: Face dataset and protocol. In *2018 international conference on biometrics (ICB)*, pages 158–165. IEEE, 2018.

[47] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14225–14234, 2021.

[48] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017.

[49] A Oppenheim, Jae Lim, Gary Kopec, and SC Pohlig. Phase in speech and pictures. In *ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 4, pages 632–637. IEEE, 1979.

[50] Alan V Oppenheim and Jae S Lim. The importance of phase in signals. *Proceedings of the IEEE*, 69(5):529–541, 1981.

[51] Nikhil R Pal and Sankar K Pal. Entropy: A new definition and its applications. *IEEE transactions on systems, man, and cybernetics*, 21(5):1260–1270, 1991.

[52] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982.

[53] 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*, pages 815–823, 2015.

[54] 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)*, pages 1–9. IEEE, 2016.

[55] Claude Elwood Shannon. A mathematical theory of communication. *ACM SIGMOBILE mobile computing and communications review*, 5(1):3–55, 2001.

[56] Abhinav Shrivastava, Abhinav Gupta, and Ross Girshick. Training region-based object detectors with online hard example mining. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 761–769, 2016.

[57] Kihyuk Sohn. Improved deep metric learning with multi-class n-pair loss objective. *Advances in neural information processing systems*, 29, 2016.

[58] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7262–7272, 2021.

[59] Yifan Sun, Changmao Cheng, Yuhan Zhang, Chi Zhang, Liang Zheng, Zhongdao Wang, and Yichen Wei. Circle loss: A unified perspective of pair similarity optimization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6398–6407, 2020.

[60] Zhenhong Sun, Ming Lin, Xiuyu Sun, Zhiyu Tan, Hao Li, and Rong Jin. Mae-det: Revisiting maximum entropy principle in zero-shot nas for efficient object detection. In *International Conference on Machine Learning*, pages 20810–20826. PMLR, 2022.

[61] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pages 10347–10357. PMLR, 2021.

[62] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 32–42, 2021.

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[64] Feng Wang, Jian Cheng, Weiyang Liu, and Haijun Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[65] 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.

[66] Xiaobo Wang, Shuo Wang, Shifeng Zhang, Tianyu Fu, Hailin Shi, and Tao Mei. Support vector guided softmax loss for face recognition. *arXiv preprint arXiv:1812.11317*, 2018.

[67] Xiaobo Wang, Shifeng Zhang, Shuo Wang, Tianyu Fu, Hailin Shi, and Tao Mei. Mis-classified vector guided softmax loss for face recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12241–12248, 2020.

[68] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, pages 499–515. Springer, 2016.

[69] Tete Xiao, Mannat Singh, Eric Mintun, Trevor Darrell, Piotr Dollár, and Ross Girshick. Early convolutions help transformers see better. *Advances in Neural Information Processing Systems*, 34:30392–30400, 2021.

[70] Mingle Xu, Sook Yoon, Alvaro Fuentes, and Dong Sun Park. A comprehensive survey of image augmentation techniques for deep learning. *Pattern Recognition*, page 109347, 2023.

[71] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generaliza-

tion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021.

[72] Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9981–9990, 2021.

[73] Yanchao Yang and Stefano Soatto. Fda: Fourier domain adaptation for semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4085–4095, 2020.

[74] Kun Yuan, Shaopeng Guo, Ziwei Liu, Aojun Zhou, Fengwei Yu, and Wei Wu. Incorporating convolution designs into visual transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 579–588, 2021.

[75] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019.

[76] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017.

[77] Xiao Zhang, Rui Zhao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. Adacos: Adaptively scaling cosine logits for effectively learning deep face representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10823–10832, 2019.

[78] Xiao Zhang, Rui Zhao, Junjie Yan, Mengya Gao, Yu Qiao, Xiaogang Wang, and Hongsheng Li. P2sgrad: Refined gradients for optimizing deep face models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9906–9914, 2019.

[79] Kai Zhao, Jingyi Xu, and Ming-Ming Cheng. Regularface: Deep face recognition via exclusive regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1144, 2019.

[80] Yaoyao Zhong and Weihong Deng. Face transformer for recognition. *arXiv preprint arXiv:2103.14803*, 2021.

[81] Zhun Zhong, Liang Zheng, Guoliang Kang, Shaozi Li, and Yi Yang. Random erasing data augmentation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 13001–13008, 2020.

[82] Benjia Zhou, Pichao Wang, Jun Wan, Yanyan Liang, and Fan Wang. Effective vision transformer training: A data-centric perspective. *arXiv preprint arXiv:2209.15006*, 2022.