FACELIVT: FACE RECOGNITION USING LINEAR VISION TRANSFORMER WITH STRUCTURAL REPARAMETERIZATION FOR MOBILE DEVICE

Anonymous Author

Anonymous Affiliation

ABSTRACT

This paper presents FaceLiVT, a lightweight yet powerful face recognition model that combines a hybrid CNN-Transformer architecture with an innovative and lightweight Multi-Head Linear Attention (MHLA) mechanism. By incorporating MHLA alongside a reparameterized token mixer, FaceLiVT effectively reduces computational complexity while preserving high accuracy. Extensive evaluations on challenging benchmarks—including LFW, CFP-FP, AgeDB-30, IJB-B, and IJB-C—highlight its superior performance compared to state-of-the-art lightweight models. The integration of MHLA significantly enhances inference speed, enabling FaceLiVT to achieve competitive accuracy with lower latency on mobile devices. Notably, FaceLiVT is 8.6× faster than EdgeFace, a recent hybrid CNN-Transformer model optimized for edge devices. With its balanced design, FaceLiVT provides a practical and efficient solution for real-time face recognition on resource-constrained platforms.

Index Terms— Face Recognition, Vision Transformer, Multi-Head Linear Attention (MHLA), Structural Reparameterization, Lightweight Model

1. INTRODUCTION

Face recognition is an important technique for identity verification in contemporary life, widely used in mobile and embedded systems for functions like device unlocking, APP access, and mobile transactions. In certain applications, such as smartphone unlocking, it requires local deployment of face verification [1]. To maintain ease of use with limited computational resources, it is crucial for mobile face verification models to be both precise and efficient. Despite advances in accuracy, cutting-edge face recognition models often use deep neural networks with numerous parameters, necessitating substantial memory and computation. Thus, deploying these advanced models on resource-limited devices, such as mobile platforms, remains a challenging task.

In order to tackle the challenges related to the memory and computational demands of cutting-edge deep neural networks, researchers are concentrating on developing lightweight and efficient neural networks for computer vision applications. These networks aim to strike a more favorable balance between recognition accuracy and the necessary memory and computational resources. Recently, there have been efforts to employ lightweight convolutional neural network (CNN) architectures, like MobileNetV1 [1], MobileNetV2 [2], and ShuffleNet [3], which have been proposed for common facial verification tasks. CNN-based models exhibit reduced model parameters and computational complexity, yet they might experience diminished accuracy levels because of the limitations in the receptive field and insufficient modeling of long-range dependencies.

Recently, transformer-based vision models have achieved remarkable success across various computer vision tasks [4, 5]. This success is largely attributed to their capacity to utilize global receptive fields and long-range dependencies, thereby outperforming CNNs in terms of performance [6]. Nevertheless, these models frequently demand considerable computational resources due to their quadratic computational complexity. For example, the original Vision Transformer (ViT) requires between 85 million and 632 million parameters for ImageNet classification. In face recognition tasks [7], ViT's computational requirements are notably high, ranging from 1.5 GFLOPs for ViT-T to 25.4 GFLOPs for ViT-L, rendering it less ideal for mobile applications [8]. Some alternative methodologies, such as EdgeFace [9], aim to decrease parameter counts and complexity by employing Low-Rank approximation with a sequence of two linear layers. However, this approach can result in slower inference speeds.

This paper introduces Linear Vision Transformer for Face Recognition (FaceLiVT) with Multi-Head Linear Attention (MHLA), an innovative and lightweight face recognition model characterized by a hybrid architecture that leverages the strengths of CNN and ViT. FaceLiVT incorporates two distinct types of token mixers: RepMix, which involves depth-wise convolution with a reparameterized kernel and residual in the initial stage, and MHLA in the final stage. MHLA involves replacing the high computation of Multi-Head Self-Attention (MHSA), thereby reducing both the number of parameters and the floating-point operations (FLOPs) required. Through comprehensive experiments on challenging benchmark face datasets—such as LFW, CFP-FP, AgeDB-30, IJB-B, and IJB-C—we demonstrate the effectiveness and efficiency of FaceLiVT compared to leading

supplementary material:link

lightweight and deep face recognition models, highlighting its suitability for deployment on resource-constrained mobile devices. The key contributions of this paper are:

1. We propose FaceLiVT, an efficient and lightweight face recognition network that combines CNN and ViT features through reparameterization, enabling real-time performance on resource-constrained platforms.

2. We introduce a Multi-Head Linear Attention (MHLA) module to reduce computational cost with linear layers while maintaining performance. MHLA replaces Self-Attention to capture spatial correlations with low complexity. To our knowledge, this is the first Hybrid CNN-Transformer incorporating MHLA for efficient face recognition.

3. We conduct extensive experiments on challenging face recognition datasets, demonstrating FaceLiVT's superior performance over existing lightweight models, along with its efficiency in mobile inference.

This paper is organized as follows: Section 2 presents a concise review of related works and Section 3 offers an in-depth description of the proposed FaceLiVT model. Section 4 outlines the experimental design. Section 5 concludes the paper and proposes directions for future research.

2. RELATED WORKS

MobileFaceNets [1, 2] represent a collection of efficient convolutional neural network models based on the MobileNetV1 and MobileNetV2 framework, designed specifically for applications in real-time face verification [2]. MobileFaceNetV1, based on the MobileNetV1 model, has achieved an accuracy of 99.4% on the LFW dataset, while MobileFaceNet, based on the MobileNetV2 architecture, achieved an accuracy of 99.7% on the LFW dataset, while maintaining a parameter count below 1 million. Inspired by ShuffleNetV2, a series of lightweight FR models termed ShuffleFaceNet was introduced in [3]. These models feature parameters ranging from 0.5 million to 4.5 million and have shown verification accuracies surpassing 99.20% on the LFW dataset.

Another approach aims to develop a face recognition model based on the original Vision Transformer (ViT) [7, 8]. This ViT achieves high accuracy on several benchmark datasets but exhibits a significant complexity of 1.5 GFLOPs, rendering it impractical for mobile applications. Based on the EdgeNeXt architecture, which serves as a hybrid model integrating the strengths of Transformers and CNNs, EdgeFace [9] is designed to decrease both the parameter count and floating-point operations (FLOPs) of this architecture, reducing parameters from 2.24 to 1.77 million and FLOPs from 196.9 to 153.9 million. This can be accomplished through low-rank linear approximation using a sequence of two linear layers, though it may result in reduced latency on mobile devices due to the use of two linear layers instead of one.

3. PROPOSED METHOD

3.1. Architecture

The prevailing macro-design approach for the FaceLiVT is largely influenced by MetaFormer [10, 11], which employs two stacked residual blocks, as depicted in Fig. 1. It initiates with a stem module that typically includes a pair of 3×3 convolutions, each with a stride of 2. From the macro perspective, the architecture incorporates a token mixer block for the extraction of spatial features, succeeded by a channel mixer block. Each block is equipped with a normalization layer and either residual or skip connections to stabilize the loss and enhance the training process. Let X_i, X_i', and $X_i'' \in \mathbb{R}^{H_i \times W_i \times C_i}$ represent the feature maps at stage i with a resolution of $H_i \times W_i$ and C_i channels with different operators; further details of the block are provided in Eq. (1) as

$$\begin{aligned} X_i' &= X_i + TokenMixer(X_i), \\ X_i'' &= X_i' + ChannelMixer(X_i'), \end{aligned} \tag{1}$$

where $TokenMixer(.)$ operator is configured as a convolution mixer or self-attention (see Table 1). $ChannelMixer(.)$ contains the Multi-Layer Perceptron (MLP) network that is conducted by two linearly fully connected layers followed by Batch Normalization (BN) and a single activation function that can be expressed in Eq. (2) as follows:

$$MLP(X_i') = BN\Big(\sigma\big(BN(X_i' * W_e)\big) * W_r\Big), \tag{2}$$

where $W_e \in \mathbb{R}^{(C_i)\times rC_i}$ and $W_r \in \mathbb{R}^{(rC_i)\times C_i}$ are the layer weights, r is the expansion ratio of the fully connected layer with a default value of 3. Operation σ is chosen using the activation function $GELU(.)$.

3.2. Structural Reparameterization

3.2.1. Fused Batch Normalization

In CNN-based facial recognition systems, convolutional layers are commonly combined with Batch Normalization (BN) layers [1, 3]. Adding BN after convolution is fundamental for improving convergence and reducing overfitting in training. However, it also elevates complexity and latency during inference. To resolve this, the BN is merged into the preceding convolution layer to form the FaceLiVT.

Convolutional layer with kernel size K, the weight matrix W is defined as $W \in \mathbb{R}^{C_o \times C_i \times K \times K}$, and the bias b as $b \in \mathbb{R}^D$, where C_i and C_o are the input and output channel dimensions, respectively. The convolution on feature $X \in \mathbb{R}^{N \times C_i \times H \times W}$ is followed by BN, involving the accumulated mean μ, accumulated standard deviation σ, feature scale γ,

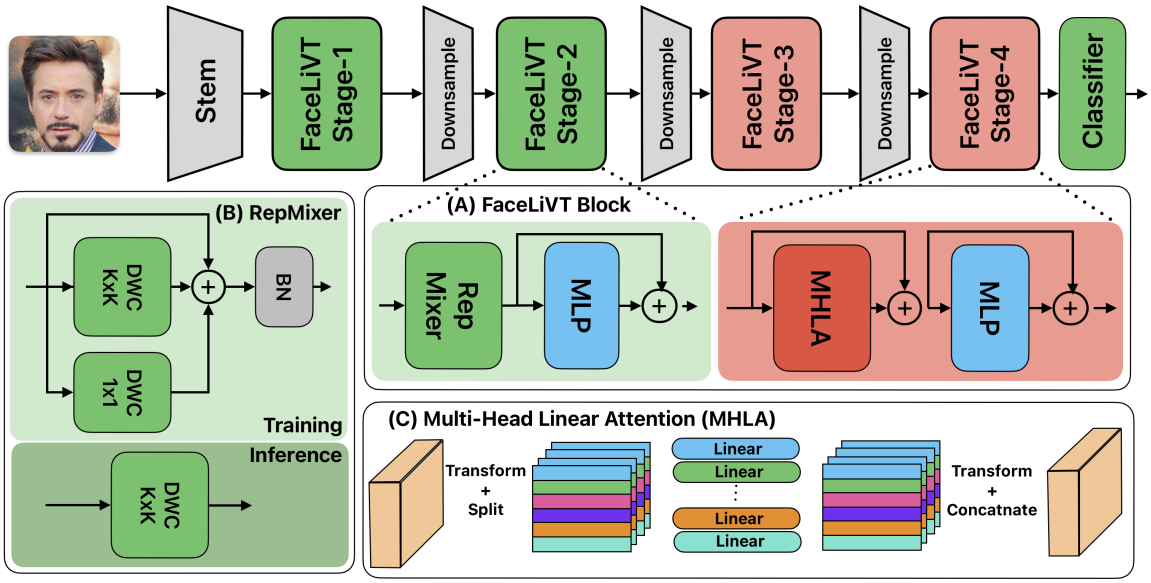

Fig. 1. FaceLiVT architecture with Multi-Head Linear Attention (MHLA) and structural reparameterization. Stages 1 and 2 use the RepMix and the last stage used MHLA as token mixer. (a) FaceLiVT Block. (b) RepMix. (c) MHLA.

bias β, and convolution operation $*$, as described in Eq.(3).

$$BN(Conv(X)) = \gamma \frac{(W * X + b) - \mu}{\sigma} + \beta. \quad (3)$$

As convolutions followed by BN during inference are linear operations, these can be merged into a single convolution layer with integrated BN, represented by Eq. (4):

$$BNConv(x) = W' \cdot X + b', \quad (4)$$

where the transformed weight is $W' = W\frac{\gamma}{\sigma}$ and the adjusted bias is $b' = (b - \mu)\frac{\gamma}{\sigma} + \beta$. BN is merged into the preceding convolutional layer across all branches, leaving only convolution in the architecture.

3.2.2. Reparameterized Token Mixer (RepMix)

The concept of convolutional mixing was initially presented in ConvMixer[12]. For an input tensor X_i, the mixing block within the layer was formulated as

$$X_i' = X_i + BN(\sigma(DWC(X_i))), \quad (5)$$

where σ denotes a non-linear activation function, and DWC is a DepthWise Convolutional layer. Although this configuration is proved quite effective, the authors of [13] modified the sequence of operations and omitted the non-linear activation. To improve the RepMix, we enhance it with a 1×1 DWC after $k \times k$ DWC, which can enhance learnability during training.

$$X_i' = X_i + \{BN(DWC_{k\times k}(X_i) + DWC_{1\times1}(X_i))\}. \quad (6)$$

To reduce the computational load and memory requirements of both the skip connection and the 1×1 DWC, these can be reparameterized into a single depthwise convolutional layer at inference time, which is especially beneficial for mobile devices.

3.3. Multi Head Self Attention

In vision transformers, the Multi-Head Self-Attention (MHSA) mechanism allows the model to evaluate token significance in a sequence for prediction and context. For an input sequence X with N tokens, MHSA computes key K, query Q, and value V through linear transformations, with $K, Q, V \in \mathbb{R}^{B\times H_e\times N\times C}$, where B is the batch size, H_e is the number of heads, N is tokens, and C is the channel dimension. Details of MHSA are outlined in Eq. (7).

$$MHSA(Q, K, V) = Concat(SA_0, .., SA_{H_e})W_o. \quad (7)$$

$$SA = Softmax\left(\frac{QK^T}{\sqrt{C}}\right)V. \quad (8)$$

where SA refers to the self-attention operation in each head. It calculates a weighted average of the values based on a similarity score between token pairs as described in Eq.8.

3.4. Multi Head Linear Attention

To reduce the computational demands while maintaining the understanding of long-range context, we present "Multi-Head Linear Attention" (MHLA). Let $X \in \mathbb{R}^{B\times C\times H\times W}$ denote the feature map with a resolution of $H \times W$ and C channels. It is first transformed to a 1D representation with N tokens, making it $X \in \mathbb{R}^{B\times C\times N}$. Subsequently, it will be divided across channels into H_e heads, resulting

in $X_{H_e} \in \mathbb{R}^{B \times \frac{C}{H_e} \times N}$. The details of MHLA are defined as follows:

$$MHLA(Q, K, V) = Concat(LA_0, .., LA_{H_e}), \quad (9)$$

where LA_{H_e} in each head comprises a sequence of two weighted linear operations with non-linear activation functions to evaluate spatial relationships among input tokens. LA_{H_e} in each head can be expressed as:

$$LA_{H_e}(X_{H_e}) = \Big(W_o\big(\sigma(X_{He} \cdot W_i)\big)\Big), \quad (10)$$

with $W_i \in \mathbb{R}^{N \times Nr}$ and $W_o \in \mathbb{R}^{Nr \times N}$ denoting the linear weights. In addition, Nr is the number of tokens with the expansion ratio r. When X_{He} is multiplied by weights W_i and W_o, the computational complexity, which depends on Nr with total complexity of $\Omega(MHLA) = 2(NNr)C$. The total complexity of MHLA is lower than the $\Omega(MHSA) = 4NC^2 + 2N^2C$ [14].

Table 1. All Variant FaceLiVT Model configurations. #Blocks denotes the number of FaceLiVT blocks.

Stage	Size	Layer	FaceLiVT			
			S	M	S-(Li)	M-(Li)
Stem	112^2	Conv	$[3 \times 3$, Stride 2$] \times 2$			
		Dims(C_i)	40	64	40	64
1	28^2	Mixer	RepMix 3×3			
		#Blocks	2	2	2	2
2	14^2	Downspl	RepMix 3×3, Stride 2			
		Dim(C_i)	80	128	80	128
		Mixer	RepMix 3×3			
		#Blocks	4	4	4	4
3	7^2	Downspl	RepMix 3×3, Stride 2			
		Dim (C_i)	160	256	160	256
		Mixer	MHSA		MHLA	
		#Blocks	6	6	6	6
4	4^2	Downspl	RepMix 3×3, Stride 2			
		Dim (C_i)	320	512	320	512
		Mixer	MHSA		MHLA	
		#Blocks	2	2	2	2
Classifier Head			Avg Pool, FC (512)			

4. EXPERIMENTS

4.1. Traning and Testing Details

FaceLiVT trained with the Glint360K dataset [17], which consists of pre-aligned 112×112 resolution facial images. They were transformed into tensors and normalized between -1 and 1. Training was performed with a batch size of 256 in each of three RTX A6000 (40GB) GPUs. AdamW optimizer with a learning rate of 6×10^{-3}, the CosFace [18] loss function, and a polynomial decay learning rate schedule were used with a 512-dimensional embedding size in the PartialFC [17] training algorithm.

We evaluated the proposed FaceLiVT model utilizing seven diverse benchmark datasets, including LFW [19], CFP-FP [20], AgeDB-30 [21], IJB-B [22], and IJB-C [23]. We provide the True Accept Rate (TAR) at a False Accept Rate (FAR) of 1e-4 for IJB-B and IJB-C datasets. In the inference speed test, the model was converted with coremltools and measured the latency on the iPhone 15 Pro.

4.2. Benchmarking Result

Table 2 provides a comparison of several face recognition models, including the FaceLiVT variants, against state-of-the-art models, with respect to parameters, computational cost (FLOP), accuracy on benchmark datasets, and mobile latency. We categorized models based on the number of FLOP around 300-1100 M FLOP and <300 M FLOP. The FaceLiVT models, which utilize a Hybrid Vision Transformer (ViT) structure, demonstrate a balance between computational efficiency and accuracy performance. In comparison with conventional CNN-based models such as MobileFaceNet and ShuffleFaceNet, the FaceLiVT models typically achieve superior accuracy on rigorous benchmarks like CFP-FP, AgeDB-30, IJB-B, and IJB-C, highlighting the advantages of the Hybrid ViT architecture in managing diverse and complex facial variations.

The integration of MHLA in the "LA" variants of FaceLiVT significantly lowers latency while preserving competitive accuracy. For instance, the FaceLiVT-M(LA) model delivers high performance on all benchmark datasets with a similar inference speed from FaceLiVT-S that used MHSA. Moreover, it can achieve competitive accuracy with 8.6 $\times$ faster than EdgeFace-XS(0.6), the recent hybrid ViT for face recognition, and 21.2 $\times$ faster than pure ViT-Based model. This underscores the capability of MHLA in enhancing computational efficacy while maintaining a satisfactory level of accuracy. It also indicates that MHLA can greatly improve real-time usability on mobile platforms. Besides its efficiency, MHLA may limit the model's capacity to capture complex long-range dependencies compared to MHSA, especially in highly unconstrained environments that lead to slight performance degradation.

4.3. Ablation Study

We conduct a 20 epoch ablation study to identify two key factors affecting the performance of FaceLiVT-S-(LA): structural reparameterization and the count of heads (H_e) MHLA mechanism. According to Table 3, structural reparameterization is pivotal in enhancing the latency of FaceLiVT-S-(LA). Eliminating reparameterization for residual and BN raises the latency from 0.47 ms to 0.50 ms and 0.60 ms, thus highlight-

Table 2. Comparison of FaceLiVT Variant with State-Of-The-Art on Face Recognition Benchmark Dataset.

Model	Type	Param (M)	FLOP (M)	Train Epoch	LFW	CFP -FP	Age DB-30	IJB		Lat (ms)
								B	C	
ViT-S [8]	ViT	86.6	5,713	40	99.8	98.9	98.3	-	96.7	14.23
TransFace-S [8]	ViT	86.7	5,824	40	99.9	98.9	98.5	-	97.3	14.31
MobileFaceNet[2]	CNN	2.0	933	20	99.7	96.9	97.6	92.8	94.7	0.77
MobileFaceNetV1[2]	CNN	3.4	1100	20	99.4	95.8	96.4	92.0	93.9	0.81
SwiftFaceFormer-L1 [15]	Hybrid	11.8	805	35	99.7	96.7	97.0	91.8	93.8	1.50
ShuffleFaceNet-1.5[2]	CNN	2.6	577	20	99.7	96.9	97.3	92.3	94.3	0.69
EdgeFace-S(0.5) [9]	Hybrid	3.6	306	50	99.8	95.8	96.9	93.6	95.6	10.21
GhoseFaceNet-V2-1 [16]	CNN	6.88	272	50	99.9	98.9	98.5	95.7	97.0	0.71
FaceLiVT-M	Hybrid	14.3	569	20	99.8	97.1	97.2	93.4	95.0	1.11
FaceLiVT-M-(LA)	Hybrid	9.75	386	20/ 40	99.7/ 99.8	96.0/ 97.2	96.7/ 97.6	92.5/ 93.7	94.1/ 95.7	0.67
ShuffleFaceNet-0.5 [2]	CNN	1.4	66.9	20	99.2	92.6	93.2	-	-	0.45
EdgeFace-XS(0.6) [9]	Hybrid	1.77	154	50	99.7	94.4	96.0	92.7	94.8	5.82
FaceLiVT-S	Hybrid	5.89	237	20	99.7	95.2	96.3	89.1	89.7	0.61
FaceLiVT-S-(LA)	Hybrid	5.05	160	20/ 40	99.6/ 99.7	94.6/ 95.1	95.6/ 96.6	83.4/ 91.2	82.5/ 92.7	0.47

ing the effectiveness of this approach in boosting computational speed. Additionally, removing $DWC_{1\times1}$ as weight refinement emphasizes its significance in accuracy around 0.9% in CFP-PP and 1.0% for Age DB-30. Importantly, these enhancements do not affect the model parameters (Param) or computational cost (FLOP), indicating that these methods refine the processing pipeline without impacting the model's overall complexity.

Table 4 shows the impact of H_e, the number of heads in MHLA, is assessed. Reducing the H_e to 8 decreases the parameter count to 4.09 M and slightly reduces FLOPs to 157 M, which enhances latency to 0.41 ms. Nevertheless, this adjustment leads to a minor decline in accuracy, with model performance registering at 99.6% on LFW and 93.9% on CFP-FP. Conversely, increasing H_e to 16 enhances accuracy to 94.6% on CFP-FP and 95.6% on AgeDB-30, accompanied by a slight rise in computational demands and latency (now 0.48 ms). This indicates that augmenting H_e bolsters the model's capability to grasp complex features at the expense of a slight decrease in runtime efficiency.

The ablation study indicates that structural reparameterization methods and the selection of H_e in MHLA are critical factors for balancing accuracy and latency in FaceLiVT-S-(LA). The best configuration, which includes fused BN, residual reparameterization, and $H_e = 16$, provides an advantageous trade-off by achieving high accuracy with minimal latency, rendering it suitable for real-time applications.

Table 3. Ablation of RepMixer block in FaceLiVT-S-(LA)

Ablation	Par (M)	FLOP (M)	LFW	CFP -FP	Age DB-30	Lat (ms)
Baseline	5.05	160	99.6	94.9	95.6	0.47
w/o Res Rep	5.05	160	99.6	94.9	95.6	0.50
w/o fused BN	5.10	160	99.6	94.9	95.6	0.60
w/o $DWC_{1\times1}$	5.05	160	99.6	93.5	94.6	0.47

Table 4. Ablation of FaceLiVT-S(LA) that using MHLA, The ablation shows the effect of number head H_e

H_e	Par (M)	FLOP (M)	LFW	CFP -FP	Age DB-30	Lat (ms)
8	4.09	157	99.6	93.9	95.0	0.41
16	5.05	160	99.6	94.6	95.6	0.48

5. CONCLUSION

The paper introduces FaceLiVT, a CNN-Transformer architecture with structural reparameterization and Multi-Head Linear Attention (MHLA) for effective face recognition on mobile platforms. Experiments on benchmarks such as LFW, AgeDB-30, CFP-FP, IJB-B, and IJB-C revealed the superior accuracy-latency balance over other lightweight models. MHLA significantly boosts inference speed while maintaining competitive performance, and reparameterization reduces computational cost without compromising accuracy. Although MHLA enhances speed and efficiency, its ability in complex long-range dependencies is less robust than full self-attention, affecting performance in environments with occlusions. Future work could explore MHLA to retain efficiency while enhancing contextual understanding.

6. REFERENCES

[1] Sheng Chen et al., "Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices," in *Bio. Recog.: 13th Chinese conf., CCBR 2018, Urumqi, China, August 11-12, 2018, Proc. 13*. Springer, 2018, pp. 428–438.

[2] Yoanna Martinez-Diaz et al., "Benchmarking lightweight face architectures on specific face recognition scenarios," *Artificial Intelligence Review*, vol. 54, no. 8, pp. 6201–6244, 2021.

[3] Yoanna Martinez-Diaz et al., "Shufflefacenet: A lightweight face architecture for efficient and highly-accurate face recognition," in *Proc. of the IEEE/CVF int. conf. on comp. vis. works.*, 2019, pp. 0–0.

[4] Enze Xie et al., "Segformer: Simple and efficient design for semantic segmentation with transformers," *Adv. in neur. inf. proc. sys.*, vol. 34, pp. 12077–12090, 2021.

[5] Nicolas Carion et al., "End-to-end object detection with transformers," in *Eur. conf. on comp. vis.* Springer, 2020, pp. 213–229.

[6] Alexey others Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *Int. Conf. on Learn. Rep. (ICLR)*, 2021.

[7] Xiang An et al., "Killing two birds with one stone: Efficient and robust training of face recognition cnns by partial fc," in *Proc. of the IEEE/CVF conf. on comp. vis. and patt. recog.*, 2022, pp. 4042–4051.

[8] Jun Dan et al., "Transface: Calibrating transformer training for face recognition from a data-centric perspective," in *Proc. of the IEEE/CVF Int. Conf. on Comp. Vis.*, 2023, pp. 20642–20653.

[9] Anjith George et al., "Edgeface: Efficient face recognition model for edge devices," *IEEE Trans. on Bio., Behav., and Id. Sci.*, 2024.

[10] Weihao Yu et al., "Metaformer is actually what you need for vis.," in *Proc. of the IEEE/CVF Conf. Comp. Vis. Patt. Recog.*, 2022, pp. 10819–10829.

[11] Weihao Yu et al., "Metaformer baselines for vis.," *IEEE Trans. on Patt. Anal. and Mach. Intell.*, 2023.

[12] Asher Trockman and J Zico Kolter, "Patches are all you need?," *Int. Conf. on Learn. Rep. (ICLR)*, 2024.

[13] Pavan Kumar Anasosalu Vasu et al., "Fastvit: A fast hybrid vis. transformer using structural reparameterization," in *Proc. of the IEEE/CVF Int. Conf. on Comp. Vis.*, 2023, pp. 5785–5795.

[14] Ze Liu et al., "Swin transformer: Hierarchical vis. transformer using shifted windows," in *Proc. of the IEEE/CVF Int. Conf. on Comp. Vis.*, 2021, pp. 10012–10022.

[15] Luis S Luevano, Yoanna Martínez-Díaz, Heydi Méndez-Vázquez, Miguel González-Mendoza, and Davide Frey, "Swiftfaceformer: An efficient and lightweight hybrid architecture for accurate face recognition applications," in *International Conference on Pattern Recognition*. Springer, 2024, pp. 244–258.

[16] Mohamad Alansari et al., "Ghostfacenets: Lightweight face recognition model from cheap operations," *IEEE Access*, vol. 11, pp. 35429–35446, 2023.

[17] Xiang An et al., "Partial fc: Training 10 million identities on a single machine," in *Proc. of the IEEE/CVF int. conf. on comp. vis.*, 2021, pp. 1445–1449.

[18] Hao Wang et al., "Cosface: Large margin cosine loss for deep face recognition," in *Proc. of the IEEE conf. on comp. vis. and patt. recog.*, 2018, pp. 5265–5274.

[19] Gary B Huang et al., "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.

[20] Soumyadip Sengupta et al., "Frontal to profile face verification in the wild," in *2016 IEEE Wint. conf. on App. of comp. vis. (WACV)*. IEEE, 2016, pp. 1–9.

[21] Stylianos Moschoglou et al., "Agedb: the first manually collected, in-the-wild age database," in *Proc. of the IEEE conf. on comp. vis. and patt. recog. workshops*, 2017, pp. 51–59.

[22] Cameron Whitelam et al., "Iarpa janus benchmark-b face dataset," in *Proc. of the IEEE conf. on comp. vis. and patt. recog. workshops*, 2017, pp. 90–98.

[23] Brianna others Maze, "Iarpa janus benchmark-c: Face dataset and protocol," in *2018 int. conf. on bio. (ICB)*. IEEE, 2018, pp. 158–165.