SwiftFaceFormer: An Efficient and Lightweight Hybrid Architecture for Accurate Face Recognition Applications

Luis S. Luevano[1], Yoanna Martínez-Díaz[3], Heydi Méndez-Vázquez[3], Miguel González-Mendoza[2], Davide Frey[1]

[1] Univ Rennes, Inria, CNRS, IRISA, France
[2]Tecnologico de Monterrey, Nuevo León, Mexico
[3]Advanced Technologies Application Center (CENATAV), Havana, Cuba

Introduction

- **Context:** Efficient Face Recognition (FR) is a key challenge in Computer Vision, prompting the exploration of strategies that **balance accuracy and computational resources**. In this work, we propose 'SwiftFaceFormer', a hybrid model combining lightweight CNNs with Vision Transformers, tailored for efficient and accurate face recognition.
- **Motivation:** Finding an adequate balance between computational efficiency and face recognition accuracy is challenging. Our work addresses finding this gap by **adapting the SwiftFormer framework specifically for face recognition**, and optimizing it for real-time performance on embedded devices.
- **Goal:** To improve and explore efficiency on Hybrid-based methods through SwiftFaceFormer. Enhance efficiency performance through strategic modifications and compensate accuracy with Knowledge Distillation, achieving a very low latency and with competitive accuracy on face recognition benchmarks. We specifically focus on efficient and accurate FR scenarios being available on **extremely hardware-constrained platforms**, such as the **Nvidia Jetson Nano**.

SwiftFaceFormer approach

- **Face recognition head**: We set the full embedding channel dimension directly after Stage 4 of the SwiftFormer architecture. We adjust the output channels the predefined face embedding dimension C using a Global DepthWise Convolution (GDC) after an efficient 1×1 point-wise convolution to reduce to the final $512\times1\times1$ embedding size. Finally, we apply another point-wise convolution and a batch normalization step.
- **SwiftFaceFormer-XXS**: We introduce a more efficient variant using an Efficient Convolutional Encoder with a decending group strategy per SwiftFormer stage and optimized depth d and width w configurations. This allows to have a strong balance between computational efficiency and limited accuracy compromises.
- **Hard Sample Knowledge Distillation**: We achieve competitive results using our XXS variant as student with our L3 variant as teacher, on popular FR benchmarks .

SwiftFaceFormer-XXS

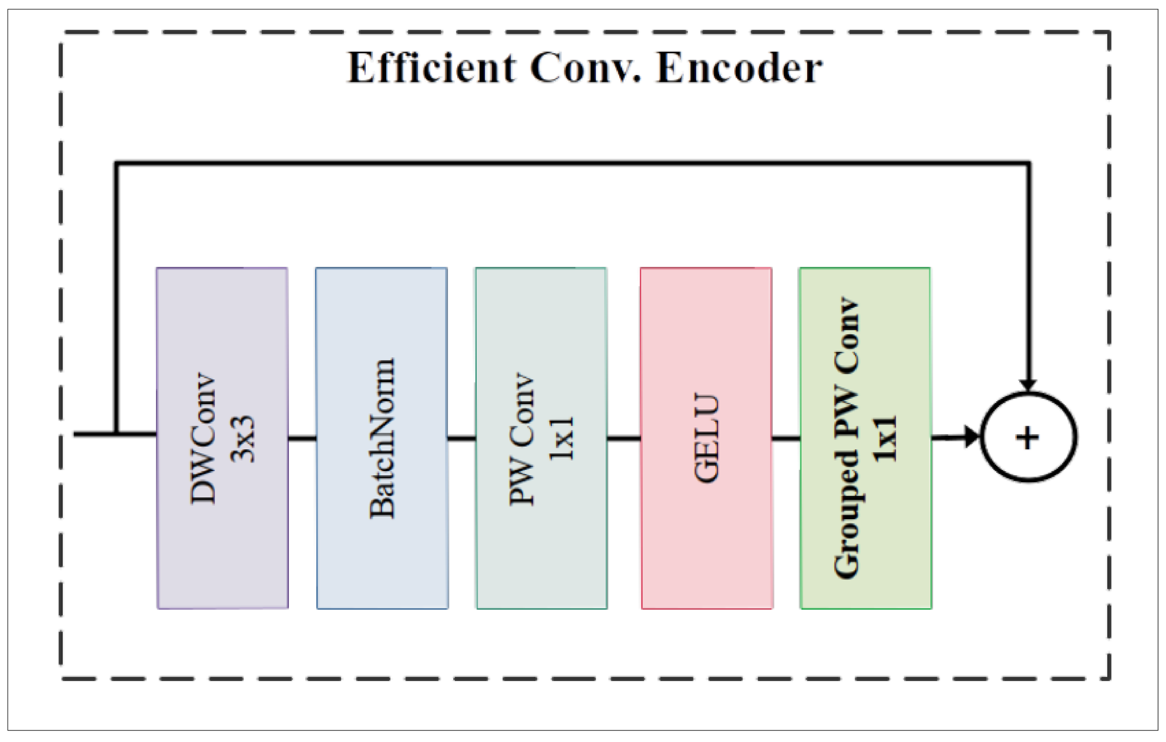

Figure: Efficient Conv. Encoder in SwiftFaceFormer-XXS. We employ Grouped Point-Wise Convolutions only at the last layer for maximizing efficiency and mitigating accuracy penalties.

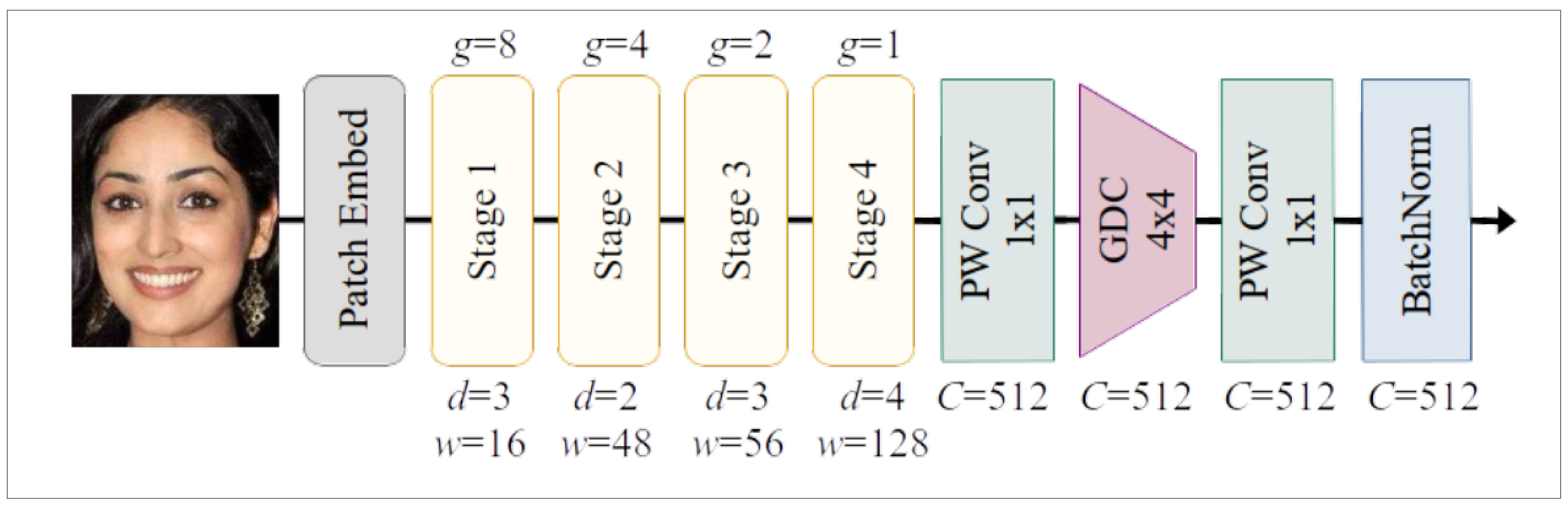

Figure: SwiftFaceFormer-XXS overall architecture. Consistent with the original SwiftFormer notation for the stages, the complexity is expressed as depth d for the number of encoding operations and width w for the number of feature map channels. C denotes the embedding channel dimension for our face recognition head.

Ablation study on CNNs and Hybrid methods

Method (%)	Params.	FLOPs (M)	LFW (M)	CFP-FP (%)	AgeDB-30 (%)	CALFW (%)	CPLFW (%)	IJB-B (%)	IJB-C (%)
MixFaceNet-M	3.9	626.1	99.68	-	97.05	-	-	91.55	93.42
MixFaceNet-S	3.1	451.7	99.60	-	96.63	-	-	90.17	92.30
ShuffleFaceNet	2.6	577.5	99.67	97.26	97.32	95.05	88.50	92.25	94.30
MobileFaceNet	2.0	933.3	99.70	96.90	97.60	95.20	89.22	92.83	94.70
PocketNetM-128-KD	1.7	1,099	99.65	95.07	96.78	95.67	90.00	90.63	92.63
MixFaceNet-XS	1.0	161.9	99.60	-	95.85	-	-	88.48	90.73
PocketNetS-128	**0.9**	587.1	99.50	93.78	95.88	95.01	88.93	88.29	90.79
PocketNetS-128-KD	**0.9**	587.1	99.55	93.82	96.50	95.15	89.13	89.23	91.47
HOTformer-Net (base)	-	1,301	99.70	**97.80**	97.60	**96.00**	91.90	**93.80**	95.50
HOTformer-Net (small)	-	765	99.70	96.50	96.90	95.60	91.10	92.50	94.50
EdgeFace-S	3.7	306.1	**99.78**	95.81	96.93	95.71	92.56	93.58	**95.63**
EdgeFace-XS	1.8	154	99.73	94.37	96.00	95.28	91.82	92.67	94.85
CFormerFaceNet	1.7	**40.0**	99.73	95.06	97.12	95.80	90.20	-	-
MobileFaceFormer	1.4	-	99.60	96.79	**97.69**	95.98	**98.43**	-	-
SwiftFaceFormer-L3	28.0	2,015.6	99.75	97.80	97.55	96.03	90.70	92.92	94.70
SwiftFaceFormer-L1	11.8	804.6	99.68	96.61	96.95	95.80	90.10	91.81	93.82
SwiftFaceFormer-S	6.0	485.2	99.60	96.49	96.83	95.78	90.00	91.56	93.54
SwiftFaceFormer-XS	3.4	293.7	99.60	95.47	96.35	95.35	88.65	90.20	92.32
SwiftFaceFormer-XXS-KD	1.5	64.1	99.43	92.50	94.82	94.78	86.97	87.81	90.28

Table: State-of-the-art CNN and Hybrid models on popular FR benchmarks for less than 4M parameters. The IJB-B and IJB-C columns correspond to the verification TAR at FAR=1e-4 on the IJB-B and IJB-C datasets, while the rest show verification accuracy (%).

Efficiency assessment

Method	Latency (ms)	FPS throughput	Params (M)	FLOPs (M)	Avg. FR Acc. (%)	Acc. per latency (%/ms)
SwiftFaceFormer-L3	36.9	27.1	28.0	2,015.6	**95.6**	2.6
SwiftFaceFormer-L1	18.0	55.3	11.8	804.6	95.0	5.3
SwiftFaceFormer-S	12.8	77.7	6.0	485.2	94.8	7.4
SwiftFaceFormer-XS	9.1	109.6	3.4	293.7	94.0	10.3
SwiftFaceFormer-XXS-KD	**4.6**	**215.5**	**1.5**	**64.1**	92.4	**20.1**

Table: Efficiency metrics tested on the Nvidia Jetson Nano platform. Our XXS-KD variant shows remarkable efficiency performance across all metrics.

Discussion

- **Efficiency performance**: We note that our SwiftFaceFormerXXS-KD exhibits the lowest latency and the highest FPS among our variants. More critically, in our 'Accuracy per latency' score, we note a huge improvement of Accuracy per latency points with the XXS-KD variant, demonstrating its feasibility for usage on real-time hardware-constrained deployments. With respect to the state of the art, this variant also outperforms the every other network in at least one efficiency metric (if more than one reported).
- **On accuracy performance**: Our SwiftFaceFormer-XS and SwiftFaceFormer-XXS-KD models, demonstrate promising results on the evaluated benchmarks, with XS obtaining as good verification results as the hybrid EdgeFace-S model and the lightweight MixFaceNet-S CNN model. Also, it is able to achieve competitive against ResNet18- Q8-bit which has more complexity requirements. KD allows us to enhance the performance of our compact SwiftFaceFormer-XXS model, offering a good trade-off between efficiency and accuracy for deploying it in limited-resource devices.

Code and Paper at:

Acknowledgements

This work was partially funded by the SOTERIA H2020 project. SOTERIA received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No101018342. This content reflects only the author's view. The European Agency is not responsible for any use that may be made of the information it contains.
The authors would like to thank the financial support from Tecnologico de Monterrey through the "Challenge-Based Research Funding Program 2022". Project ID # E120- EIC-GI06 - B-T3 - D