Facial Attribute Recognition using Lightweight Multi-Label CNN-Transformer Architecture for Intelligent Advertising

Adri Priadana
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
priadana3202@mail.ulsan.ac.kr

Muhamad Dwisnanto Putro
Department of Electrical Engineering
Universitas Sam Ratulangi
Manado, Indonesia
dwisnantoputro@unsrat.ac.id

Jinsu An
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
jinsu5023@islab.ulsan.ac.kr

Duy-Linh Nguyen
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
ndlinh301@mail.ulsan.ac.kr

Xuan-Thuy Vo
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
xthuy@islab.ulsan.ac.kr

Kang-Hyun Jo
Department of Electrical, Electronic and Computer Engineering
University of Ulsan
Ulsan, Korea
acejo@ulsan.ac.kr

***Abstract*—In modern cities, intelligent advertising platforms have been widely engaged in public areas. A facial attribute recognition technique is essential to assist these platforms in delivering suitable adverts for each audience. These platforms also require a recognition technology that can operate at least suitably on a CPU device to reduce implementation costs. This work proposed a lightweight multi-label CNN-Transformer architecture with an efficient inception block (EIB) and squeeze channel transformer encoder (SCTE) to perform facial attribute recognition efficiently. EIB is used to extract face features in multi-scale and levels supported by SCTE in improving its feature map's quality. The proposed architecture produces fewer parameters with low operations and gains competitive accuracy on the CelebA and LWFA datasets consisting of images with multi-label. Moreover, the proposed architecture integrated with face detection can perform sufficiently on a CPU configuration in real-time with 21 frames per second (FPS) using 224×224 input size of face area image.**

***Index Terms*—Convolutional Neural Network (CNN), Facial Attribute Recognition, Multi-label CNN Transformer, Self-Attention Module**

I. Introduction

Advertising is an essential sector in accelerating economic and social development. Advertising can increase sales, build brand awareness, provide a competitive advantage, reach and engage potential customers, and be a cost-effective marketing channel for businesses. With the advancement of artificial intelligence and information technology, intelligent advertising communication methods are constantly evolving. Digital signage, as one of the advertising platform, has become progressively across-the-board and appear in public areas in modern cities [1], [2], including department store, airport, tourism attraction, etc. This platform's presence can boost retail activity and elicit positive consumer responses [3]. As a new advertising technique employing multimedia screens board, digital signage can deliver the dynamic customization of promotional content [4]. However, this platform still suffers from presenting targeted advertising and promotion, i.e., providing promotional content to audiences who may be interested in the advertised products, leading to expending a not optimal budget [5].

In recent years, facial attributes analysis has drawn much interest from the computer vision community due to its widespread applications, including face recognition [6], face image synthesis [7], and face retrieval [8]. Moreover, this technology can also be used to support an advertising task. For example, the advertising platform can recognize the attributes used by or attached to the audience facing the platform, such as heavy makeup, bushy eyebrows, bald head, wearing a hat, eyeglasses, lipstick, etc. Thus, the platform can provide advertising related to these attributes. This mechanism can make the advertising more targeted. Even though the facial attributes recognition task is merely an image-level classification problem, it is challenging because of the low speed of the classification process and the variety of facial appearances brought on by notable variations in viewpoint, illumination, etc [9].

The majority of state-of-the-art facial attribute classification techniques utilize Convolutional Neural Networks (CNNs) to classify facial attributes due to CNNs' exceptional effectiveness. Liu et al. [10] applied a pair of CNNs (LNets+ANet) for face localization and facial attribute recognition. Hand and Chellappa [11] proposed a multi-task CNN used to divide attribute classifiers into groups, integrated with an auxiliary

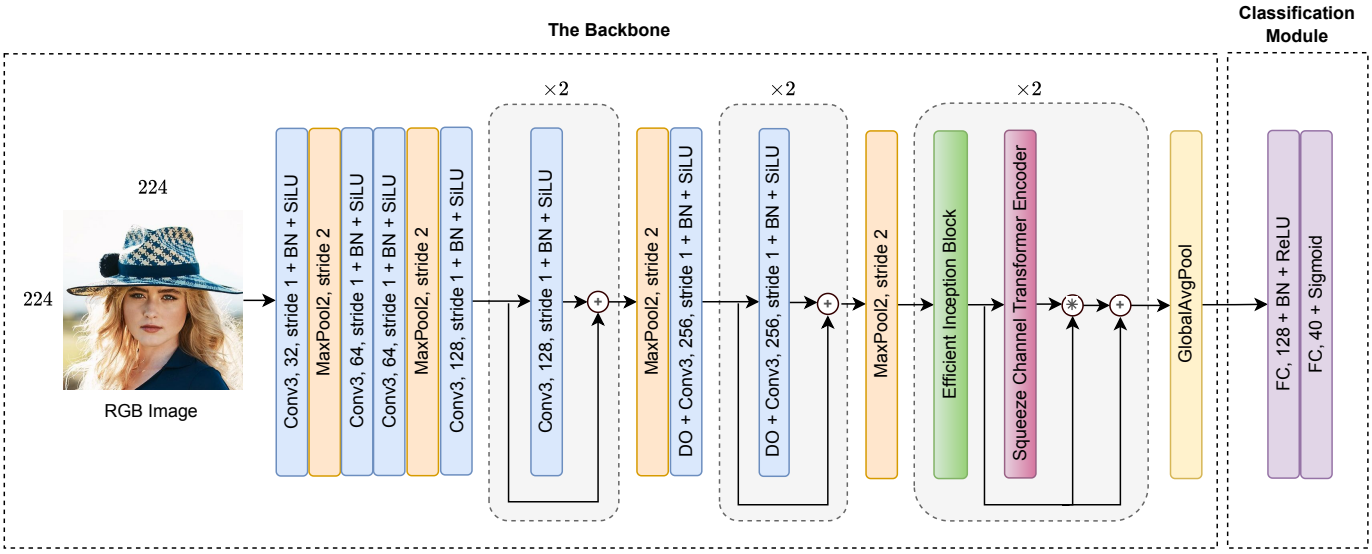

Fig. 1. The proposed facial attribute recognition architecture.

network (MCNN-AUX) that utilized the relationships among the facial attributes to improve classification performance. Novel multi-task learning of cascaded CNN called MCFA [12] was offered to predict numerous facial attributes.

A deep CNN-based method, SPLITFACE [13], was also introduced to predict facial attributes, especially in partially occluded face scenarios. Mao et al. [9] proposed a novel deep multi-task multi-label CNN, dubbed DMM-CNN, used to operate facial attribute classification. The multi-task mechanism was designed to extract features for two groups of attributes, objective and subjective. Recently, a multi-zone transformer based on self-distillation (MZTS) [14] is also offered for facial attribute classification. The last two mentioned works provide the highest performance. However, they also bring heavy parameters and computation, leading to unsuitable implementation on low-cost or CPU devices. In the actual application of digital signage, facial attribute recognition is required to run on low-cost devices to push down implementation costs [15], [16]. This platform demands an efficient recognition architecture suitable for operating on low-cost or CPU devices in real-time while maintaining its performance.

A lightweight multi-label CNN-Transformer architecture with an efficient inception block (EIB) and squeeze channel transformer encoder (SCTE) is proposed in this work. The efficient inception block employs multi-kernel-size and multi-level in an efficient manner to extract face features in high-level feature maps. The squeeze spatial transformer encoder is used to improve the feature map's quality. As a result, the architecture can perform facial attributes accurately and efficiently. Here is a summary of the main contributions of this work:

1) A lightweight multi-label CNN-Transformer architecture with soft computation and generating low parameters is offered to recognize facial attributes applied to support intelligent advertising. This architecture gains very competitive accuracy compared with other architectures on two datasets, CelebA [17] and LFWA [18].
2) An efficient inception block (EIB) is proposed to espouse the backbone extracting the facial features efficiently. The EIB applies multi-kernel-size and multi-level convolution layers to capture different scale areas and levels to enhance the variety of the feature map efficiently.
3) A squeeze channel transformer encoder (SCTE) is also presented as an enhancement module to acquire the spatial relationship representations of the features map. It efficiently stimulates the feature map quality, enhancing the recognition performance.

II. Proposed Architecture

The proposed lightweight multi-label CNN-Transformer architecture for facial attribute recognition generates 2,111,292 parameters. This architecture consists of backbone and classification modules, as illustrated in Fig. 1.

A. The Backbone

The backbone module, used to extract facial features from a face, engages nine of 3×3 convolution layers sequentially, which grows from 32 to 256. Batch Normalization (BN) [19] is applied in every convolution layer, followed by Sigmoid Linear Unit (SiLU) activation [20], to deal with the gradient issue. This backbone also involves a shortcut connections technique as a residual mapping [21] in the last two convolution layers of 128 and 256 channels to construct the output. Due to the sequential convolution layers, this backbone applies four times downsampling using 2×2 max-pooling operations with strides 2. Applying only nine convolution layers causes the architecture is not deep enough to extract the facial features. Therefore, we propose an efficient inception block (EIB) as an additional extractor and a squeeze channel transformer encoder (SCTE)

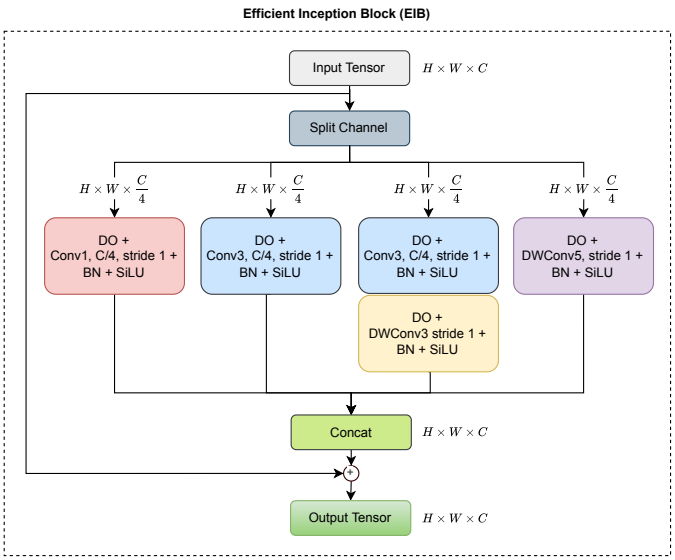

Fig. 2. The proposed Efficient Inception Block (EIB).

to generate spatial relationship representations efficiently. Both parts are placed and performed two times after the last max-pooling layer and before the global average-pooling operation.

B. Efficient Inception Block (EIB)

Motivated from the inception block [22] to go deeper in extracting the image features, this module applies four branches of the convolution layer shown in Fig. 2. Unlike the original inception block that uses the same number of channels as an input, the proposed efficient inception block splits the input feature map $\mathbf{X}$ into four elements $[\mathbf{X_1}, \mathbf{X_2}, \mathbf{X_3}, \mathbf{X_4}]$ based on channel axes, then performs and combines convolution operations with various kernel sizes and levels. To reduce the number of parameters, we apply a depthwise convolution layer in the 5×5 and second level of 3×3 branch formulated as follows:

$$\begin{aligned} EIB(\mathbf{X}) = \mathbf{X} + Concat[&\delta(BN(C1(DO(\mathbf{X_1})))), \\ &\delta(BN(C3(DO(\mathbf{X_2})))), \\ &\delta(BN(DWC3(DO(\delta(BN(C3(DO(\mathbf{X_3}))))))))), \\ &\delta(BN(DWC5(DO(\mathbf{X_4}))))], \end{aligned} \tag{1}$$

where $C1$, $C3$, $DWC3$, and $DWC5$ are convolution layers with 1×1, 3×3 kernel sizes, depth-wise convolution layers with 3×3, and 5×5 kernel sizes, respectively. BN and DO indicate batch normalization and dropout operations, respectively. δ is Scaled Sigmoid Linear Unit (SiLU) activation, and $Concat$ is the concatenate operation.

C. Squeeze Channel Transformer Encoder (SCTE)

Following the victorious use of the Vision Transformer (ViT) technique [23] inspired by the Transformer model [24], modern networks focus on enhancing the self-attention mechanism. However, the overhead of the transformer self-attention approach cannot be applied optimally for vision tasks on low-cost devices, since it results in greater multiplication and addition operations. This work proposes a Squeeze Channel Transformer Encoder (SCTE) to relieve the high operations problem. This encoder first squeezes the feature map in spatial axes by aggregating spatial information across the channel using 1×1 convolution operation with Rectified Linear Unit (ReLU) activation to produce a thin query, key, and value, as shown in Fig. 3. This operation will compute a tensor input $\mathbf{X}$ of shape $H \times W \times C$ into a query ($\mathbf{Q}$), key ($\mathbf{K}$), and value ($\mathbf{V}$) projections with shape $H \times W \times 1$, followed by reshaping operation to produce a $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ matrix with shape $H \times W$. After that, we compute these matrices using scaled dot-product attention (SDPA) represented as follows:

$$SDPA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = softmax\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \tag{2}$$

where T is a transpose matrix operation and d_k is a scaling factor to control the softmax temperature. The second reshaping operation is also performed to restore the output matrix shape into $H \times W \times 1$. In this encoder, the SDPA will compute the relationship between each other of spatial information rows. We also apply the multi-head attention mechanism to perform Squeeze Channel Multi-Head Attention (SCMHA). This mechanism linearly projects the $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ head (h) times with distinct, learned linear projections using 1×1 convolution operation with Rectified Linear Unit (ReLU) activation, followed by SDPA in a parallel process. All output from all heads are concatenated and once again projected using 1×1 convolution operation with a dropout (DO) and Rectified Linear Unit (ReLU) activation, resulting in the final tensor. At the last stage, we perform a linear projection using 1×1 convolution operation with a Rectified Linear Unit (ReLU) activation. We utilize a residual connection [21] around the SCMHA and last convolution layer.

D. Classification Module

This classification module is employed to classify facial features extracted by the backbone to compute the probability of each facial attribute class. This module has two fully connected (FC) layers with 128 and 40 units, respectively. The first FC layer applies Batch Normalization (BN) and Rectified Linear Unit (ReLU) activation, while the last FC layer uses only Sigmoid activation. The Sigmoid activation will convert the input into the independent probability score. Therefore, multiple labels potentially have a great score independently, which means the instances belonging to multiple labels or classes. Then, the Cosine Similarity loss is used to compare the prediction result vector with the ground truth vector.

III. Implementation Setup

The training process of the proposed architecture in this work is performed on NVIDIA GTX1080Ti 11GB, using CelebA and LFWA datasets implemented on Tensorflow and Keras framework. It utilizes a data augmentation strategy, applying rotation, rescaling, shifting, shearing, zooming, and

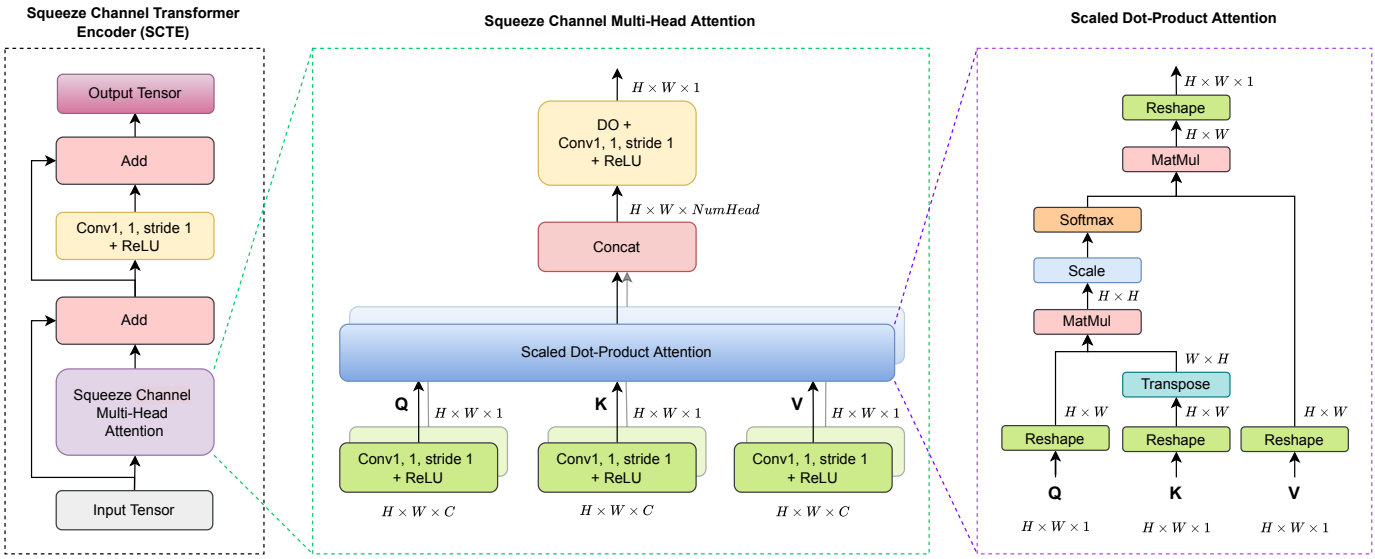

Fig. 3. The proposed Squeeze Channel Transformer Encoder (SCTE).

horizontal flipping to enhance knowledge and prevent overfitting problems. The initial learning rate is set to 10^{-3} and will decrease to 75% when there is no revision in 5 epochs. The Adam optimizer is used to rectify the weight based on the Cosine Similarity loss. The d_k value in SCTE is empirically set to be 2. The proposed architecture is trained with a batch size of 32 in 30 and 60 epochs for CelebA and LFWA datasets, respectively. Moreover, an Intel Core i7-9750H CPU @ 2.60GHz with 20GB RAM is utilized to test the proposed architecture in the real-time scenario.

IV. Experimental Results

A. Evaluation on Datasets

1) CelebA: CelebFaces Attributes (CelebA) dataset consists of 202,599 face images with multi-label covering high pose variations and background clutter labeled with 40 binary attributes. This dataset provides the aligned and cropped images version. Following the previous works [9], [14], this dataset is split into 162,770 for training, 19,867 for validation, and 19,962 for testing. By employing only 2,111,292 parameters, the proposed architecture gains 91.50% in average accuracy and placed third best, which differed by 0.2% and 0.16% with the best and the second best, respectively, as seen in Table I. Nevertheless, the proposed architecture has far fewer parameters. The detailed accuracy for each facial attribute of the CelebA dataset is shown in Fig. 4.

2) LFWA: Labeled Faces in the Wild Attributes (LFWA) dataset consists of 13,143 face images with multi-label covering different lighting, poses, ages, occlusions, and expressions. This dataset is labeled with 73 binary attributes. Following the previous works [9], [14], we select the same 40 attributes from LFWA as CelebA. This dataset is split into 6,572 for training and 6,571 for testing. The proposed architecture gains 86.45% in average accuracy, and placed third best, which differed by 0.28% and 0.11% with the best and the second best, respectively, as seen in Table II. Even so, the proposed architecture generates distant fewer parameters. The detailed accuracy for each facial attribute of the LFWA dataset is also shown in Fig. 4.

TABLE I
The evaluation results on CelebA dataset.

Architectures	Input Size (Pixel)	Data Augmentation	Number of Parameters (Million)	Average Accuracy (%)
PANDA [25]	64 × 64	No	-	85.43
LNets+ANet [10]	227 × 227	Yes	100	87.33
SPLITFACE [13]	196 × 196	Yes	26.09	90.61
MOON [26]	178 × 218	No	119.7	90.94
MCFA [12]	224 × 224	No	260	91.23
SOP [27]	224 × 224	No	4.99	91.26
MCN-AUX [11]	224 × 224	No	16	91.29
MZTS [14]	224 × 224	No	85.83	**91.66**
DMM-CNN [9]	224 × 224	No	360	**91.70**
Proposed	**224 × 224**	**Yes**	**2.11**	**91.50**

TABLE II
The evaluation results on LFWA dataset.

Architectures	Input Size (Pixel)	Data Augmentation	Number of Parameters (Million)	Average Accuracy (%)
PANDA [25]	64 × 64	No	-	81.03
LNets+ANet [10]	227 × 227	Yes	100	83.85
MCFA [12]	224 × 224	No	260	83.63
SPLITFACE [13]	196 × 196	Yes	26.09	85.82
MCN-AUX [11]	224 × 224	Yes	16	86.31
DMM-CNN [9]	224 × 224	No	360	**86.53**
MZTS [14]	224 × 224	No	85.83	**86.73**
Proposed	**224 × 224**	**Yes**	**2.11**	**86.45**

B. Ablation Study

1) Model Analysis: This ablation study is conducted on the CelebA dataset by revoking the module and then computing the average accuracy to investigate the effect of each component in the proposed architecture. We also investigate

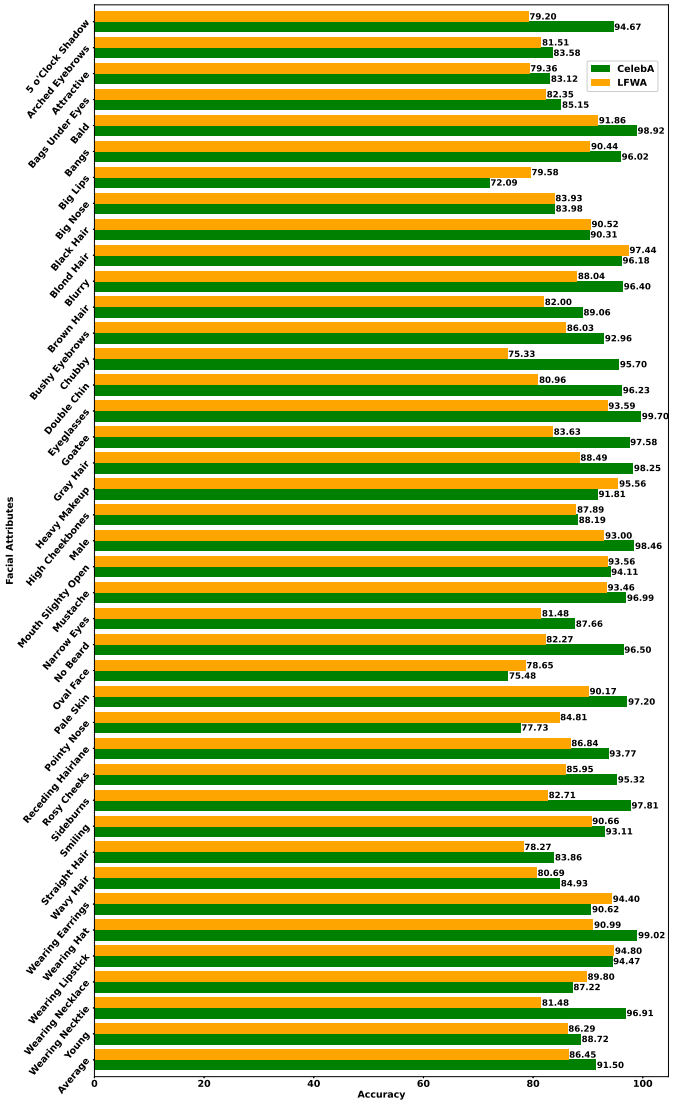

Fig. 4. Detail accuracy for each facial attribute on CelebA and LFWA datasets.

TABLE III
THE MODEL ANALYSIS ON CELEBA DATASET.

Data Augmentation	EIB	SCTE	MFLOPs	Number of Parameters	Average Validation Accuracy	Average Testing Accuracy (%)
			6,123	1,944,488	91.92	79.93
✓			6,123	1,944,488	91.96	91.37
✓	✓		6,186	2,107,688	91.98	91.42
✓	✓	✓	6,188	2,111,292	91.96	91.50

the impact of the data augmentation strategy on recognition performance. As can be seen in Table III, performing the data augmentation strategy to overcome the overfitting problem can increase the average accuracy significantly by 11.44%. Moreover, the proposed EIB and SCTE, with only generating more few parameters, can enhance the average accuracy by 0.05% and 0.08%, respectively.

2) Number of Head Analysis: This analysis is conducted by setting the different numbers of heads on the SCTE and then computing the average accuracy to investigate the optimal number of attention heads. In the proposed architecture, using a single head does not improve performance, as shown in Table IV. On the other side, too many numbers of heads also do not provide optimal average accuracy in this scenario. The proposed SCTE with two heads provides the highest average accuracy in this work.

TABLE IV
NUMBER OF HEAD ANALYSIS ON CELEBA DATASET.

Number of Heads	Number of Parameters	MFLOPs	Average Accuracy (%)
1	2,109,748	6,187	91.42
2	**2,111,292**	**6,188**	**91.50**
3	2,112,836	6,188	91.48
4	2,114,380	6,189	91.47
5	2,115,924	6,190	91.43

TABLE V
RUNTIME EFFICIENCY WITH DIFFERENT INPUT SIZE ON INTEL CORE I7-9750H CPU @ 2.60GHZ WITH 20GB RAM.

Input Size (Pixel)	Number of Parameters	MFLOPs	Average Accuracy (%)	FAR (FPS)	FAR + FD (FPS)
224×224	2,111,292	6,188	91.50	23.14	21.16
112×112	2,111,292	1,547	91.00	72.57	56.56
56×56	2,111,292	386	90.08	143.37	92.61

FAR indicates the Facial Attribute Recognition
FAR + FD indicates the Facial Attribute Recognition integrated with Face Detection

C. Runtime Efficiency

The practical application of digital signage requires facial attribute recognition, integrated with face detection, to run on a CPU device in real-time to suppress implementation costs. In this scenario, we investigate the runtime of the proposed architecture in three different input sizes, 224×224, 112×112, and 56×56 pixel. Table V indicates that the smaller input size will generate lower FLOPs, leading to performing faster recognition. However, it will decrease the average accuracy because a smaller size of input images will acquire less information. We also integrate the proposed face attribute recognition with an efficient face detector named LWFCPU [28] that generates few parameters. The proposed face attribute recognition architecture will use the Region of Interest (ROI) of the face, which has been extended to cover the entire head and neck area, which comes from the face detection operation, as an input. It will be cropped and resized to a specific size appropriate for the input size image of the proposed face attribute recognition architecture. As a result, the proposed architecture can operate 23.14 frames per second (FPS) in recognizing face attributes of the human face and 21.16 FPS when integrated with face detection using 224×224 input size of face area image. This result indicates that the proposed architecture is sufficient to implement on a CPU device to support intelligent advertising platforms, such as digital signage.

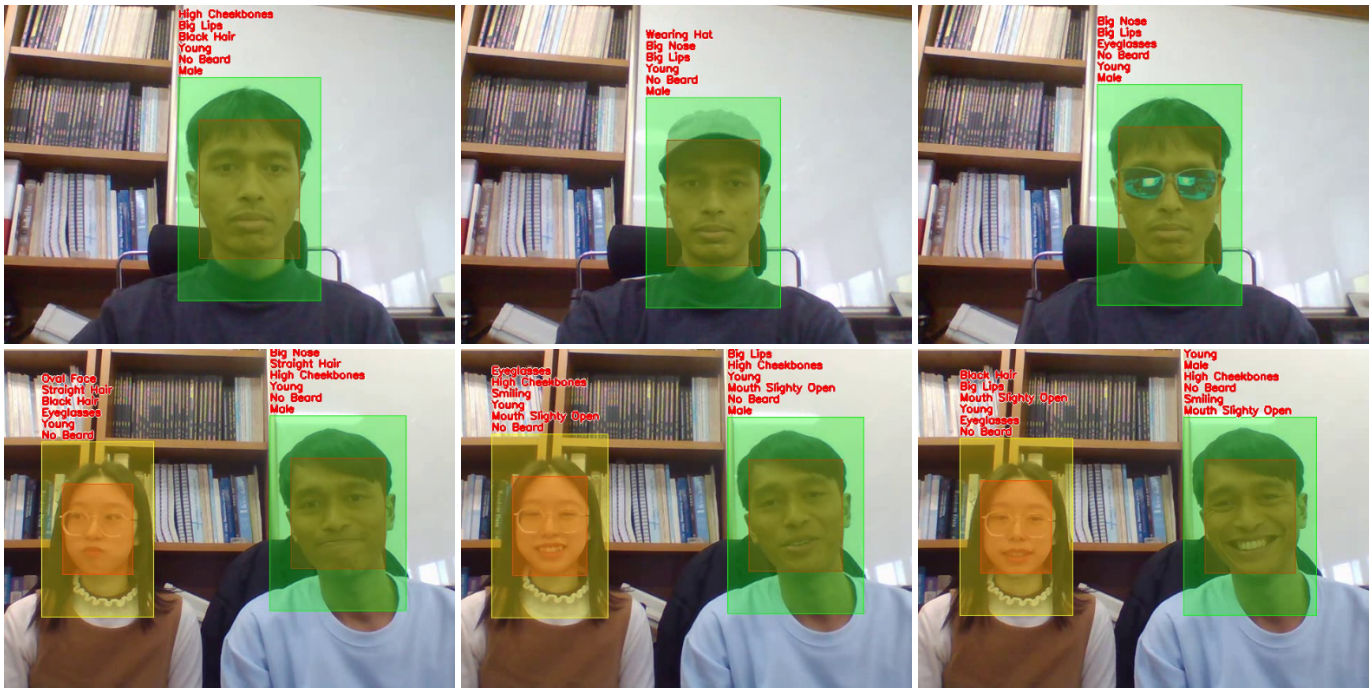

Fig. 5. Some examples of facial attribute recognition results from the proposed recognizer. It shows the top six facial attributes on the screen based on the highest probability value of the label or class, sorted from the highest value from the bottom (the attribute that appears in the lowest position on the screen has the highest probability value).

D. Qualitative Results and Discussion

Fig. 5 displays the facial attribute recognition result of the proposed architecture on the CelebA dataset, integrated with face detection. This scenario only shows the top six facial attributes on the screen based on the highest probability value of the label or class, sorted from the highest value from the bottom (the attribute that appears in the lowest position on the screen has the highest probability value). The proposed recognizer can recognize facial attributes from the face such as eyeglasses, wearing a hat, smiling, young, straight hair, black hair, high cheekbones, no beard, and even gender. Even though the CelebA dataset only has a male gender label and does not have a female gender label, it is possible to recognize which face is female based on the probability value of the male gender label. If the value of the male class is extremely low or even close to zero, it reflects that the detected face is the opposite of the male, which is female. In Fig. 5, the green bounding box of the face represents a male, while the yellow one indicates a female. The red bounding box describes the face ROI derived from the face detection operation.

V. Conclusion

This work proposes facial attribute recognition using a lightweight multi-label CNN-Transformer architecture with low operation. This work offers an efficient inception block (EIB) and squeezes channel transformer encoder (SCTE) to help the architecture rapidly extract various facial features and enhance their quality. The proposed architecture gained competitive performance compared to the state-of-the-art on CelebA and LFWA datasets. Moreover, the proposed architecture can perform sufficiently on a CPU configuration in real-time with 23 FPS in recognizing face attributes of the human face and 21 FPS when integrated with face detection using 224×224 input size of face area image. In future work, other methods will be sought to make recognition architecture more efficient and perform faster on a CPU device, especially with 224×224 input size of face area image that provides higher accuracy than using the smaller size of the input image.

References

[1] X. Zhang, X. Xie, Y. Wang, X. Zhang, D. Jiang, C. Yu, and Y. Liang, "A digital signage audience classification model based on the huff model and backpropagation neural network," *IEEE Access*, vol. 8, pp. 71 708–71 720, 2020.

[2] A. Priadana, M. D. Putro, and K.-H. Jo, "An efficient face gender detector on a cpu with multi-perspective convolution," in *2022 13th Asian Control Conference (ASCC)*. IEEE, 2022, pp. 453–458.

[3] M. Garaus, U. Wagner, and R. C. Rainer, "Emotional targeting using digital signage systems and facial recognition at the point-of-sale," *Journal of Business Research*, vol. 131, pp. 747–762, 2021.

[4] A. Greco, A. Saggese, and M. Vento, "Digital signage by real-time gender recognition from face images," in *2020 IEEE International Workshop on Metrology for Industry 4.0 & IoT*. IEEE, 2020, pp. 309–313.

[5] L. Wang, Z. Yu, D. Yang, H. Ma, and H. Sheng, "Efficiently targeted billboard advertising using crowdsensing vehicle trajectory data," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 2, pp. 1058–1066, 2020.

[6] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper, "A comprehensive study on face recognition biases beyond demographics," *IEEE Transactions on Technology and Society*, vol. 3, no. 1, pp. 16–30, 2022.

[7] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, "3d aided duet gans for multi-view face image synthesis," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 8, pp. 2028–2042, 2019.

[8] H. M. Nguyen, N. Q. Ly, and T. T. Phung, "Large-scale face image retrieval system at attribute level based on facial attribute ontology and deep neuron network," in *Intelligent Information and Database Systems: 10th Asian Conference, ACIIDS 2018, Dong Hoi City, Vietnam, March 19-21, 2018, Proceedings, Part II*. Springer, 2018, pp. 539–549.

[9] L. Mao, Y. Yan, J.-H. Xue, and H. Wang, "Deep multi-task multi-label cnn for effective facial attribute classification," *IEEE Transactions on Affective Computing*, vol. 13, no. 2, pp. 818–828, 2020.

[10] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 3730–3738.

[11] E. Hand and R. Chellappa, "Attributes for improved attributes: A multi-task network utilizing implicit and explicit relationships for facial attribute classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 31, no. 1, 2017.

[12] N. Zhuang, Y. Yan, S. Chen, and H. Wang, "Multi-task learning of cascaded cnn for facial attribute classification," in *2018 24th International Conference on Pattern Recognition (ICPR)*. IEEE, 2018, pp. 2069–2074.

[13] U. Mahbub, S. Sarkar, and R. Chellappa, "Segment-based methods for facial attribute detection from partial faces," *IEEE Transactions on Affective Computing*, vol. 11, no. 4, pp. 601–613, 2018.

[14] S. Chen, X. Zhu, D.-H. Wang, S. Zhu, and Y. Wu, "Multi-zone transformer based on self-distillation for facial attribute recognition," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2023, pp. 1–7.

[15] K. Mishima, T. Sakurada, and Y. Hagiwara, "Low-cost managed digital signage system with signage device using small-sized and low-cost information device," in *2017 14th IEEE Annual Consumer Communications & Networking Conference (CCNC)*. IEEE, 2017, pp. 573–575.

[16] A. Priadana, M. D. Putro, X.-T. Vo, and K.-H. Jo, "A facial gender detector on cpu using multi-dilated convolution with attention modules," in *2022 22nd International Conference on Control, Automation and Systems (ICCAS)*. IEEE, 2022, pp. 190–195.

[17] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[18] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, October 2007.

[19] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[20] S. Elfwing, E. Uchibe, and K. Doya, "Sigmoid-weighted linear units for neural network function approximation in reinforcement learning," *Neural Networks*, vol. 107, pp. 3–11, 2018.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[22] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2818–2826.

[23] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[24] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[25] N. Zhang, M. Paluri, M. Ranzato, T. Darrell, and L. Bourdev, "Panda: Pose aligned networks for deep attribute modeling," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1637–1644.

[26] E. M. Rudd, M. Günther, and T. E. Boult, "Moon: A mixed objective optimization network for the recognition of facial attributes," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*. Springer, 2016, pp. 19–35.

[27] Y. Lu, A. Kumar, S. Zhai, Y. Cheng, T. Javidi, and R. Feris, "Fully-adaptive feature sharing in multi-task networks with applications in person attribute classification," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1131–1140.

[28] M. D. Putro, D.-L. Nguyen, and K.-H. Jo, "Lightweight convolutional neural network for real-time face detector on cpu supporting interaction of service robot," in *2020 13th International Conference on Human System Interaction (HSI)*. IEEE, 2020, pp. 94–99.