Ethnicity Classification Based on Facial Images using Deep Learning Approach

Abdul-aziz Kalkatawi, Usman Saeed
Dept. of Computer Science and Artificial Intelligence-College of Computer Science and Engineering,
University of Jeddah, Jeddah, Saudi Arabia

***Abstract*—Race and ethnicity are terminologies used to describe and categorize humans into groups based on biological and sociological criteria. One of these criteria is the physical appearance such as facial traits which are explicitly represented by a person's facial structure. The field of computer science has mostly been concerned with the automatic detection of human ethnicity using computer vision-based techniques, where it can be challenging due to the ambiguity and complexity on how an ethnic class can be implicitly inferred from the facial traits in terms of quantitative and conceptual models. The current techniques for ethnicity recognition in the field of computer vision are based on encoded facial feature descriptors or Convolutional Neural Network (CNN) based feature extractors. However, deep learning techniques developed for image-based classification can provide a better end to end solution for ethnicity recognition. This paper is a first attempt to utilize a deep learning-based technique called vision transformer to recognize the ethnicity of a person using real world facial images. The implementation of Multi-Axis Vision Transformer achieves 77.2% classification accuracy for the ethnic groups of Asian, Black, Indian, Latino Hispanic, Middle Eastern, and White.**

***Keywords*—Vision transformer; deep learning; ethnicity; race; classification; recognition**

I. Introduction

The terms race and ethnicity are often used interchangeably which leads to misconception in some circumstances. The word race is used to categorize humans into groups biologically based on physical appearance traits inherited from the ancestors [1], whereas the term ethnicity is used to categorize humans into groups ethnographically based on geographic regions, language, cultural tradition, and shared ancestry which could refer to the similar physical appearance traits inherited but not inclusively [2].

Racial categories were first proposed in 1779s by Johann Friedrich Blumenbach, these categories were Ethiopian-black race, Caucasian-white race, Mongolian-yellow race, American-red race, and Malayan-brown race [3]. A commonly adopted racial categorization is proposed by the U.S. Census Bureau where they categorize race into White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian and Pacific Islander [4]. The White race represents the ethnic groups originating in Europe, the Middle East, and North Africa. The Black race represents the ethnic groups originating in South Africa, Nigeria, Ghana, Kenya, etc. The American Indian or Alaska Native race represents the ethnic groups originating in North and South America also including Central America. The Asian race represents the ethnic groups originating in East or Southeast Asia, and the Indian subcontinent. The Native Hawaiian and Pacific Islander race represent the ethnic groups originating in Hawaii, Guam, Samoa, and other Pacific Islands [5].

The human face conveys a set of semantic traits; these traits can be used to conclude several attributes for a person such as identity, gender, age, race or ethnicity, and expressions [6]. The human face is the area from the upper edge of the forehead to the chin and from the left ear to the right ear. The structure of the facial area is represented in three main regions which are superior, middle, and inferior. The superior region describes the shape of the forehead, eyebrows, and eyes. The middle region describes the shape of the nose, cheeks, and ears. The inferior region describes the shape of the lips, chin, and jawline [7]. Thereby, the shape of the facial structure provides discriminant appearance traits from one person to another's. In facial recognition systems based on computer vision techniques the shape of the facial structure is referred to as facial features. The complexity of facial recognition systems lies in the process of transformation from visual facial features to a quantitative representation of the data.

Majority of the proposed methods are based on facial features descriptors where pre-defined procedures are performed to capture and analyze facial images to construct a geometry map of facial traits such as the shapes of the mouth, nose, eyes and facial landmarks or image texture such as skin color. Then the extracted features are encoded into a feature vector to be used in a classifier [8]. However, recent methods are mainly based on the automation of feature extraction using deep learning such as convolutional neural network (CNN) models, which have achieved better accuracy and generalization results when trained with a sufficient amount of representative data [8].

The lack of exploitation of deep learning techniques other than CNN motivated the study of deep learning techniques that can model the facial features for ethnicity recognition. This paper employs the deep learning model Multi-axis Vision Transformer (MaxViT) proposed by Google research team [9] for the purpose of image-based classification. The objective is to test the capability of MaxViT to recognize cognate facial features that implicitly represent the discriminative appearance traits which distinguish one ethnic group from other using facial images. The proposed model is trained on a database created by merging three different ethnicity datasets namely FairFace [10], UTKFace [11], and Arab face dataset [12]. The main contribution is that the proposed model achieves better generalization capabilities compared to other models with an

accuracy of 77.2% for classifying six ethnic groups i.e., Asian, Black, Indian, Latino Hispanic, Middle Eastern, and White. The utilization of deep learning techniques such as MaxViT would significantly improve the current state of the art for ethnicity recognition with implication for various fields such as human computer interaction and video surveillance.

II. Related Work

A. Databases

One of the crucial factors for the advancement in the scope of race or ethnicity recognition in computer vision is the availability of a large and diverse dataset that provides reliable annotated facial images based on racial or ethnic categories. However, the research area of ethnicity recognition is still lacking in this factor, as no dataset that represents all the racial or ethnic groups is available. One of the most recently proposed datasets is called FairFace [10] consisting of 97,698 images for seven ethnic groups (Black, East Asian, Indian, Latino Hispanic, Middle Eastern, Southeast Asian, and White) labeled by age, gender, and ethnicity. Another dataset proposed in the field of ethnicity recognition by Zhifei Zhang et al. [11] called the UTKFace dataset consists of 20,000 images for five ethnic groups (Asian, Black, Indian, White, Others) labeled by age, gender, and ethnicity. The dataset proposed by Ziwei Liu et al. [13] called Labelled Faces in the Wild (LFW) consists of 13,233 images for three ethnic groups (Asian, Black, White) labeled by gender, and ethnicity. MORPH dataset proposed by Karl Ricanek et al. [14] consists of 55,134 images for five ethnic groups (African, European, Asian, Hispanic, Others). BUPT-BALANCEDFACE dataset proposed by Mei Wang et al. [15] consists of more than one million images for four ethnic groups (Asian, African, Indian, and Caucasian). BUPT-GLOBALFACE dataset proposed by Mei Wang et al. [16] consists of two million images for four ethnic groups (Asian, African, Indian, and Caucasian). Mivia Ethnicity Recognition (VMER) dataset composed from VGG-Face2 dataset [8,17] and consisting of more than three million images for the ethnic groups (African American, East Asian, Caucasian Latin, and Asian Indian). There are many other facial datasets such as Diversity in Faces (Dif) [18], IMDB-WIKI dataset [19], and Cross-Age Reference Coding (CARC) dataset [20], these datasets are not optimally oriented toward ethnicity recognition.

B. Conventional Feature Extraction

This section summarizes the methods that have been commonly used for facial features extraction using computer-vision techniques for race or ethnicity recognition.

A study conducted by L. Farkas [21] which is based on the relations between well-defined facial landmarks in terms of the Euclidean distance between two points, the angle formed by a point and two other points, and the perpendicular distance from a point to the straight line between two other points. This study shows that these relations can be used to distinguish the differences in facial features of different ethnic groups. Therefore, the use of geometric facial features to classify ethnic groups is applicable. On the other hand, Xiaoguang et al. [22] used appearance-based approaches that extract facial features based on the pixel intensity values in a black-and-white image of the face. This method achieved high accuracy when implemented to classify between two ethnic groups Non-Asian, and Asian, however, this method may be insufficient to classify between more specific ethnic groups because it can vary significantly based on images quality factors such as resolution, viewing angles, and illumination. Another appearance-based approach proposed by G. Zhang et al. [23] which is based on the invariant of monotonic transformation in the grayscale images using Local Binary Pattern histograms to describe the texture and shape variations. S. Hosoi et al. [24] extracted ethnic facial features using Gabor Wavelet Transformation besides Retina sampling. Their proposed method achieved a high accuracy relative to the number of ethnic groups concluded in the experiment. An approach proposed by N. Narang et al. [25] to extract facial features from images by locating eye centers using manual annotation and affine transformation to construct a geometric representation of face images. Kazimov T. et al. [26] proposed a method to define ethnic features based on the Euclidean distance between 30 geometric landmarks. H. Ding et al. [27] proposed an approach based on 3D face models where ethnic features are extended using Oriented Gradient Maps. M. A. Uddin et al. [28] proposed an integrated approach to classify the ethnicity of Caucasian, African, and Asian based on texture and shape features using a histogram of oriented gradients and Gabor filter to extract features from a grayscale image, and then combining both feature vectors into one.

C. Deep Learning-based Feature Extraction

This section summarizes the deep learning approaches that have been proposed previously for feature extraction for ethnicity recognition.

Marwa Obayya et al. [29] used a fusion of three pre-trained CNN models as feature extractors, namely VGG16, Inception v3, and capsule networks. And a bidirectional long short-term memory model as a classifier, the model is trained using VMER dataset and achieves an accuracy of 70% for classifying four ethnic groups of African American, East Asian, Caucasian Latin, and Asian Indian. Gurram Sunitha, K. et al. [30] used a pre-trained Xception CNN model as a feature extractor and kernel extreme learning machine model is used as classifier. The model is trained using the BUPT-GLOBALFACE dataset and achieves an accuracy of 97% for classifying four ethnic groups of Asian, African, Caucasians, and Indian. Norah A. Al-Humaidan et al. [12] used a pre-trained ResNet50 CNN model as a feature extractor and a fully connected layer for classification. The model is trained on a sub-ethnic group of Arabs dataset consisting of 5,598 images of Gulf Cooperation Council (GCC) countries people, 1,665 images of Levant people, and 1,555 images of Egyptian people. The model achieved 76% classification accuracy. Heng Zhao et al. [31] proposed ethnicity recognition framework by utilizing a CNN model, Content-Based Image Retrieval model (CBIR), and Support Vector Machines (SVM) classifier. A VGG-16 CNN model is used for feature extraction and a Bag-of-Words model is used as CBIR, a combination of CNN feature and ranking feature are used to train SVM model for classification. The model is trained using a dataset consisting of 1,000 images of Bangladeshi people, 1,520 images of Chinese people, and 1,078 images of Indian people. The model achieved 95%

classification accuracy. Hu Han et al. [32] used a modified AlexNet CNN model with batch normalization layers for feature extraction and two fully connected layers for classification. The model is trained on MORPH-II dataset achieving 96% classifying accuracy for three ethnic groups of Black, White, and Other. Anwar Inzamam et al. [33] used a pre-trained VGG-Face CNN model for feature extraction and a SVM model as a classifier. The model is trained on ten different databases, using ten-fold cross-validation where nine databases are used for training and one for testing. The model achieved 98% average classification accuracy over all databases for three ethnic groups of Asian, White, and Black. Amr Ahmed et al. [34] used a Feed-Forward based CNN model and max pooling layer for classification. The model is trained using Face Recognition Grand Challenge dataset achieving 93% classifying accuracy for three ethnic groups of Asian, White, and Other.

A summary of related work based on deep learning approach is shown in Table I, describing the model used, the number of ethnicity groups classified by the model and accuracy achieved by the model compared to the proposed model.

TABLE I. A SUMMARY OF RELATED WORK BASED ON DEEP LEARNING APPROACH

Author	Method	Ethnicity groups	Accuracy
Marwa Obayya et al. [29]	Fusion of VGG16, Inception v3, and capsule network CNN models	African American, East Asian, Caucasian Latin, and Asian Indian	70%
Gurram Sunitha, K. et al. [30]	Xception CNN model	Asian, African, Caucasians, and Indian	97%
Norah A. Al-Humaidan et al. [12]	ResNet50 CNN model	GCC people, Levant, and Egyptian	76%
Heng Zhao et al. [31]	VGG-16 CNN model	Bangladeshi, Chinese, and Indian	95%
Hu Han et al. [32]	AlexNet CNN model	Black, White, and Other	96%
Anwar Inzamam et al. [33]	VGG-Face CNN model	Asian, White, and Black	98%
Ahmed et al. [34]	Feed-Forward based CNN model	Asian, White, and Other	93%
Proposed model	MaxVit vison transformer model	Asian, Black, Indian, Latino Hispanic, Middle Eastern, and White	77%

III. PROPOSED METHOD

This section describes the proposed approach for ethnicity recognition using computer-vision techniques based on deep learning. There are two main limitations of the existing techniques described in the literature review section. First, most of the proposed techniques are limited to classifying up to four ethnic groups. Secondly, the proposed techniques are limited to the utilization of CNN models for the purpose of feature extraction. Thus, this paper proposes a model for classifying six ethnic groups i.e., Asian, Black, Indian, Latino Hispanic, Middle Eastern, and White by employing the MaxViT model which is a hybrid Vision Transformer based model capable of feature extraction and classification.

A. Multi-Axis Vision Transformer (MaxViT)

Initially transformers were proposed for the task of natural language processing [35], the prime feature of transformers is the self-attention mechanism which is the ability to capture semantic relations between data segments in a sequence. However, recently in the scope of computer-vison transformers have attracted considerable interest in the research community and several approaches have been proposed for image classification, segmentation, object detection, and generation. Thereby, Zhengzhong Tu et al. [9] proposed a Self-Attention mechanism named multi-axis self-attention (Max-SA) which can capture both local and global semantic relations between data segments. This is accomplished by decomposing the self-attention mechanism into window attention for local interaction and grid attention for global interaction. The Max-SA mechanism is a stand-alone attention module which can be adopted in any network architecture. Thus, The Max-SA module is the backbone structure of MaxViT model (see Fig. 1) coupled with Inverted Residual Block (MBConv) [36]. The model is available on Google Colab notebook (https://colab.research.google.com/drive/1UvseIP7zvFiysagSp4zfvt9f9ErHu-lo?usp=sharing).

MaxViT module uses the relative positional multi-head attention mechanism. The basic concept of an attention mechanism is to estimate the relevance of one data token to other data tokens in a sequence. In self-attention layer there are three trainable weight matrices (W^Q, W^K, W^V) from which three variables are generated by performing dot-product multiplication of the initial input variable (X_i) with learnable matrices represented as $(Q = XW^Q, K = XW^K, V = XW^V)$, from which attention layer output is represented as in Eq. (1), where d_k is input size [37].

$$Attention(Q, K, V) = sofmax\left(\frac{QK^t}{d_k}\right)V \quad (1)$$

As for multi-head attention which is an extension of self-attention where the input is first partitioned into several segments and each segment is processed in parallel by a separate attention layer from which the output of each layer is considered as an attention head. Hence, multiple attention heads are aggregated as the final output allowing the model to capture various feature aspects of the input. As for the relative positional self-attention, an additional bias is concatenated with the output of the attention layer which incorporates positional importance of data tokens in a sequence.

The MaxViT module is composed of three main blocks. First, the MBConv with Squeeze-and-Excitation (SE) [38] block, window attention block, and grid attention block. The MBConv with SE is utilized to enhance the model efficiency

and generalization, where MBConv are used to scale the model depth wise allowing it to capture complex features, and SE are used as channels wise self-attention mechanism that capture interdependencies between channels. The window attention block transforms the input feature map into non-overlapping windows to represent a confined attention by reshaping it $\left(\frac{H}{P} \times \frac{W}{P}, P \times P, C\right)$ where P is the window size. The grid attention block transforms the input feature map into uniform grid to represent a sparse attention by reshaping it $\left(G \times G, \frac{H}{G} \times \frac{W}{G}, C\right)$ where G is the grid size. Each of the attention blocks outputs are reshaped back to the initial input shape and passed through a multi-Layer perceptron block is illustrated in Fig. 1.

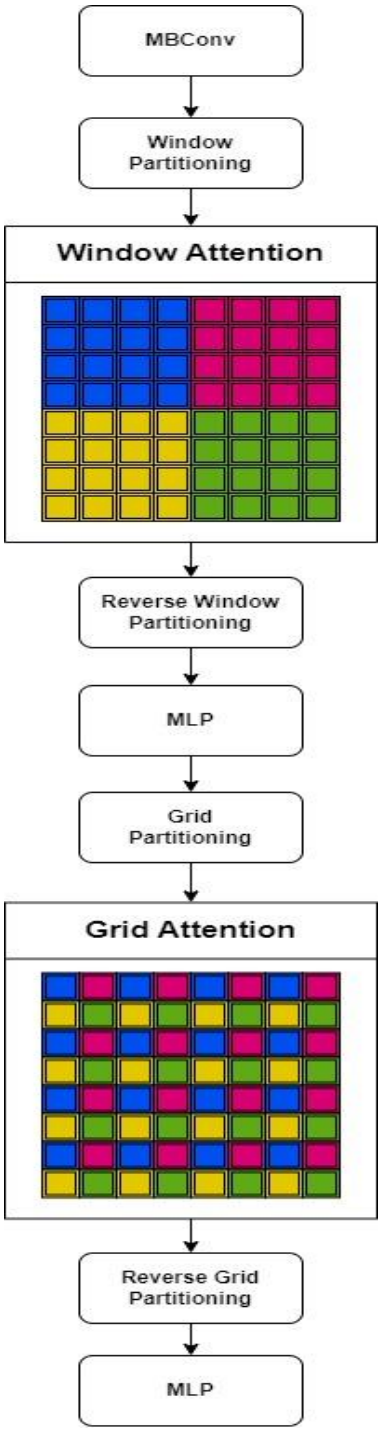

Fig. 1. MaxViT module attention mechanism.

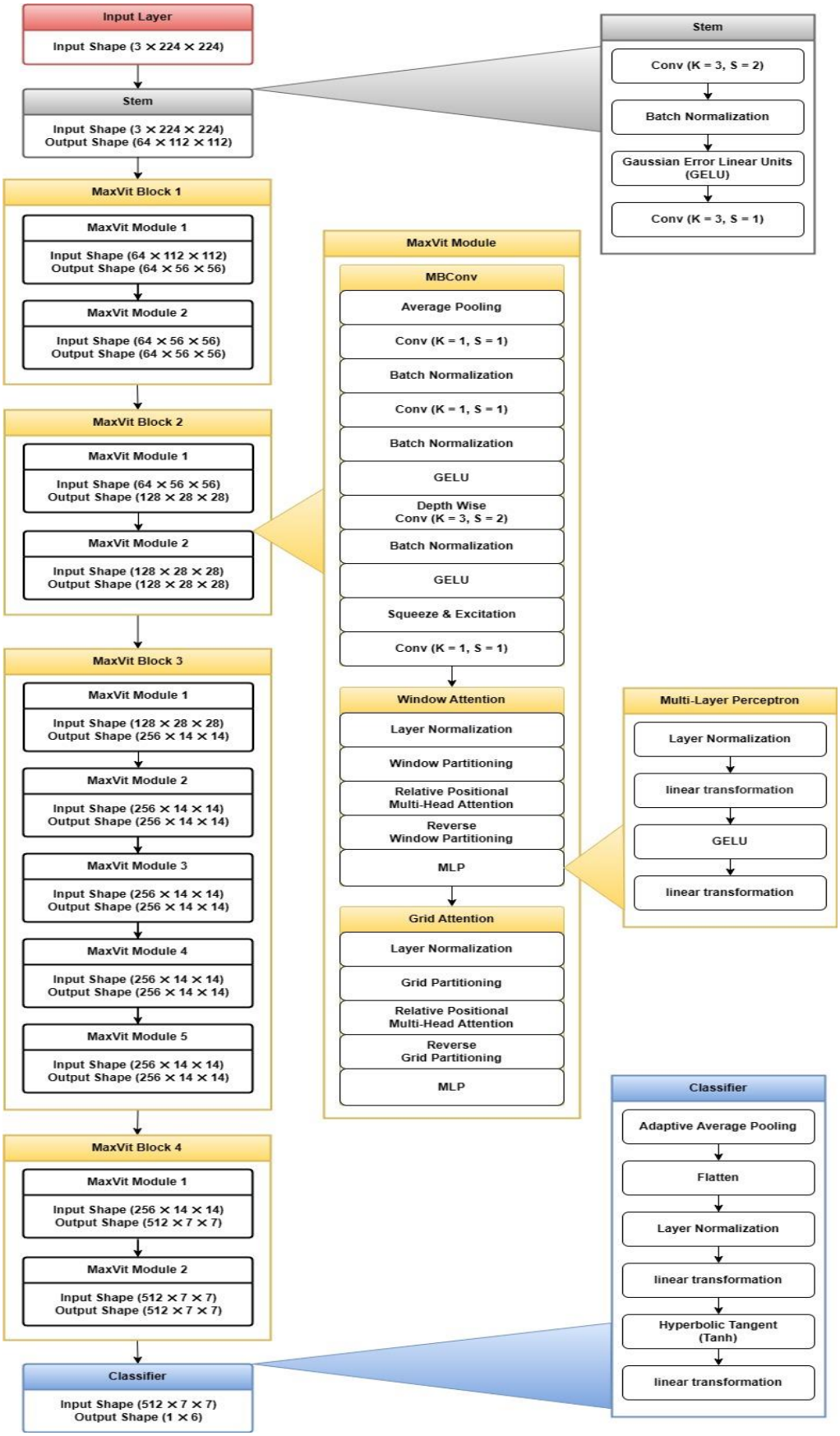

Fig. 2. Architecture of MaxVit model.

The MaxViT model architecture is shown in Fig. 2. The MaxViT model architecture can be described as follows. First, the input layer which takes an input feature map of size (C, H, W) where C depicts the feature map channels/depth, H the feature map height, and W the feature map width. The stem layer uses convolutional layers to extract low-level features from the input and reduce its spatial dimensionality. Thus it reduces the computational complexity of the model. The MaxVit block which is composed of sequentially staked MaxViT modules where each block outputs half the resolution of the prior block with a doubled channels size. Finally, the classifier which transforms the multi-dimensional output into one-dimensional i.e., a feature vector. From then the feature vector is passed to a fully connected layer which performs linear transformation and outputs a prediction.

IV. Experiment and Results

This section describes the datasets used for model training and testing, the experiment conducted, and the results obtained. The experiments were implemented using PyTorch (2.0.0+cu118) for Python (3.10.5) and executed on a computer with Intel Core i7-6700 processor with 16 GB RAM, and RTX 3080 with 10-GB VRAM GPU.

A. Dataset

The experiments were conducted on a database created by merging three datasets FairFace [10], UTKFace [11], and Arab face dataset [12]. Dataset is described in Fig. 3 composed of six classes with sample sizes of 15,937 for Asian, 18,589 for Black, 18,074 for Indian, 14,988 for Latino Hispanic, 15,188 for Middle Eastern and 28,645 White, with a total of 111,421 samples split into 101,474 samples for training and 9,947 samples for testing. The ratio of training samples to testing samples per class is shown in Fig. 4. Random samples from each dataset are shown in Fig. 5.

B. Experiment

This section describes the configuration and hyperparameters used for the proposed model. The objective of this experiment is to employ the MaxVit transformer-based model for ethnicity recognition using facial images. The experiment utilizes transfer learning technique to reduce computational complexity and training time by reusing the pre-trained parameters of all the model layers excluding the classifier head which were modified to an output size of 6 hence the initial model is trained on ImageNet dataset [39] for object classification with an output size of 1000. Also, in the experiments all of the pre-trained model layers parameters are retrained.

Data Preprocessing: Typically, when utilizing transfer learning data must follow the same preprocessing pipeline used for training the initial model. Thus, for data preprocessing the first step is to resize the image to 224 × 224 pixels with center crop applied, and then a random horizontal flip of the image is applied with probability of 80% as a data augmentation. Next, image pixels values are converted from 0 to 255 to be between 0.0 and 1.0, where each of the color channels values represented as (Red, Green, Blue) are normalized by a mean of 0.485, 0.456, 0.406 and standard deviation of 0.229, 0.224, 0.225, which can enhance the model's learning process by standardizing the input. The specific values are often determined empirically based on the dataset being used. In this case, these values are used when trained on ImageNet dataset.

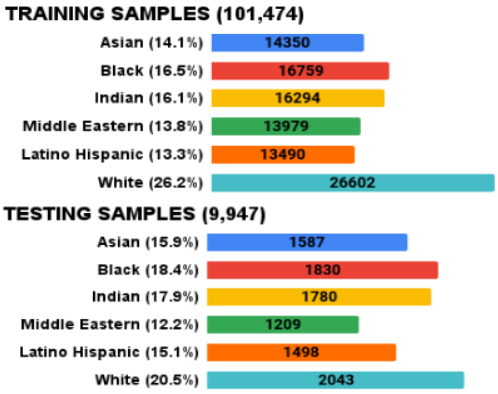

Fig. 3. Dataset overview.

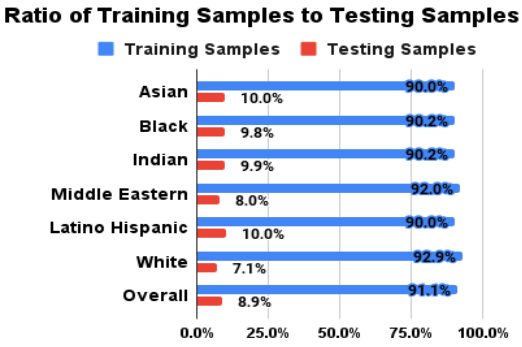

Fig. 4. Data split ratio.

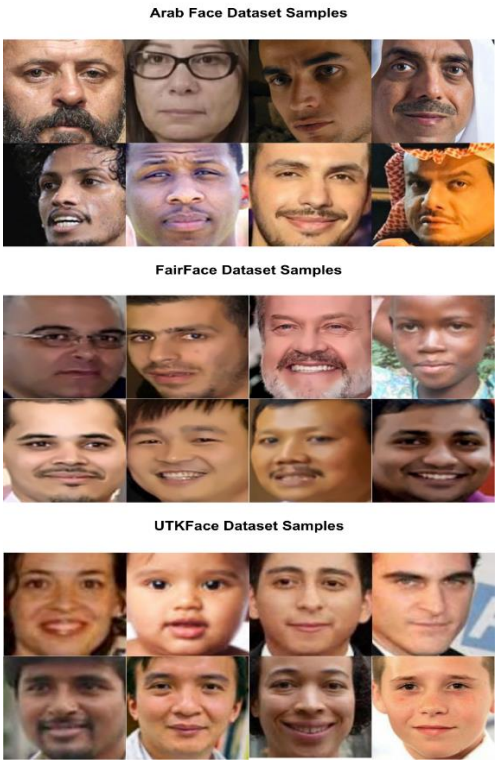

Fig. 5. Samples overview from each dataset.

Model hyperparameters: The model takes an input tensor shape of (B, C, H, W) where B stands for batch size and has the value of 20 images, C for color channels which is 3 for each image, H for the pixel height of each image that is 224, and W for the pixel width of each image that is 224. Based on the experiment an optimal batch size is between 16 and 20, thus for the purpose of reducing training time and fully utilizing hardware capacity the batch size is set as 20 images. A head dimension of 32 is used to represent the output feature map of the attention layers with partitioning size of 7 × 7 for both window and grid attentions, for the purpose of reusing the pre-trained parameters. Cross-entropy loss is used as a loss function to measure the dissimilarity between predicted probabilities and the actual targets. As for model parameters optimization the Adadelta algorithm is implemented with a learning rate of 0.1. Adadelta is an adaptive learning rate technique [40] which dynamically and automatically adjusts the learning rates on a per-parameter level, thus based on the experiment with Adadelta optimizer the learning rate value does no substantial impact on the learning process unless extreme values are used.

C. Results

In the main experiment the model was trained for 15 epochs achieving the highest classification accuracy of 0.772 for classifying six ethnic groups of Asian, Black, Indian, Latino Hispanic, Middle Eastern, and White. Additional experiments are conducted for a comparison in which four CNN models are trained using the same dataset. These models are a pre-trained VGG-Face model based on Vgg-16 architecture [41] which is developed for face recognition with over two million faces images. A pre-trained VGG-Face2 model based on ResNet-50 architecture [17] which is also developed for face recognition but with over three million faces images. A pre-trained EfficientNet-V2 model [42] for object classification is also based on MBConv. Additionally for the purpose of testing MaxVit model scalability an experiment is conducted by training the proposed model on three ethnic classes i.e., Black, White and others which is a merged class of all the remaining categories. For training the sample sizes used are 17,015 samples for Black, 17,099 samples for White, and 18,458 samples for the merged class. For evaluation the sample sizes used are 1,300 samples for black, 1,300 samples for white and 1,500 samples for the merged class. The model achieved a classification accuracy of 0.835. Lastly for comparison the same experiment is conducted with three classes using the AlexNet model [32] which achieved the classification accuracy of 0.782. The results are shown in Fig. 6. Additionally, the classification accuracy of top two predicted classes is shown in Fig. 7 where the proposed MaxVit model achieves the highest score of 91.3%. A comparison of model size in terms of parameters size is shown in Fig. 8, where the proposed MaxVit model being the smallest model in terms of parameters size. Hence, in terms of performance smaller models require lower computational capacity thus being more efficient in terms of speed and size on disk. Confusion matrices are shown in Fig. 9 describing the models classification performance of 9,947 samples for six classes.

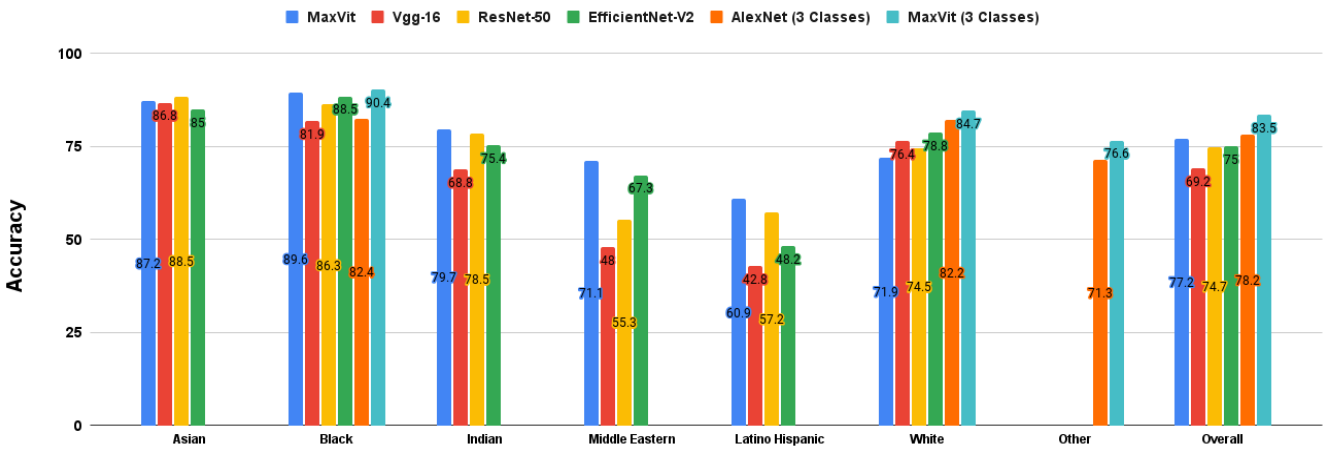

Fig. 6. Models classification accuracies.

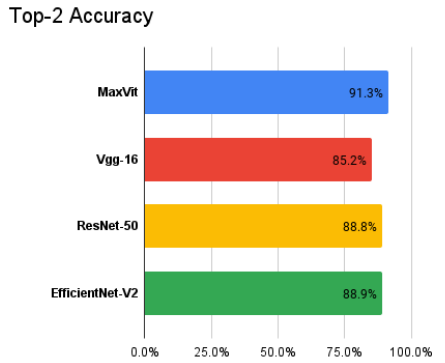

Fig. 7. Top two predicted classes accuracy scores.

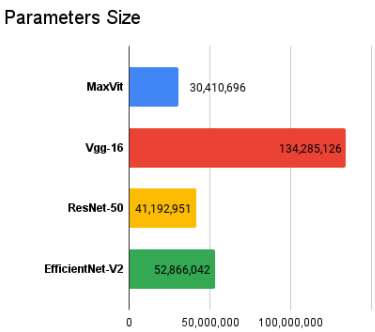

Fig. 8. Models parameters size.

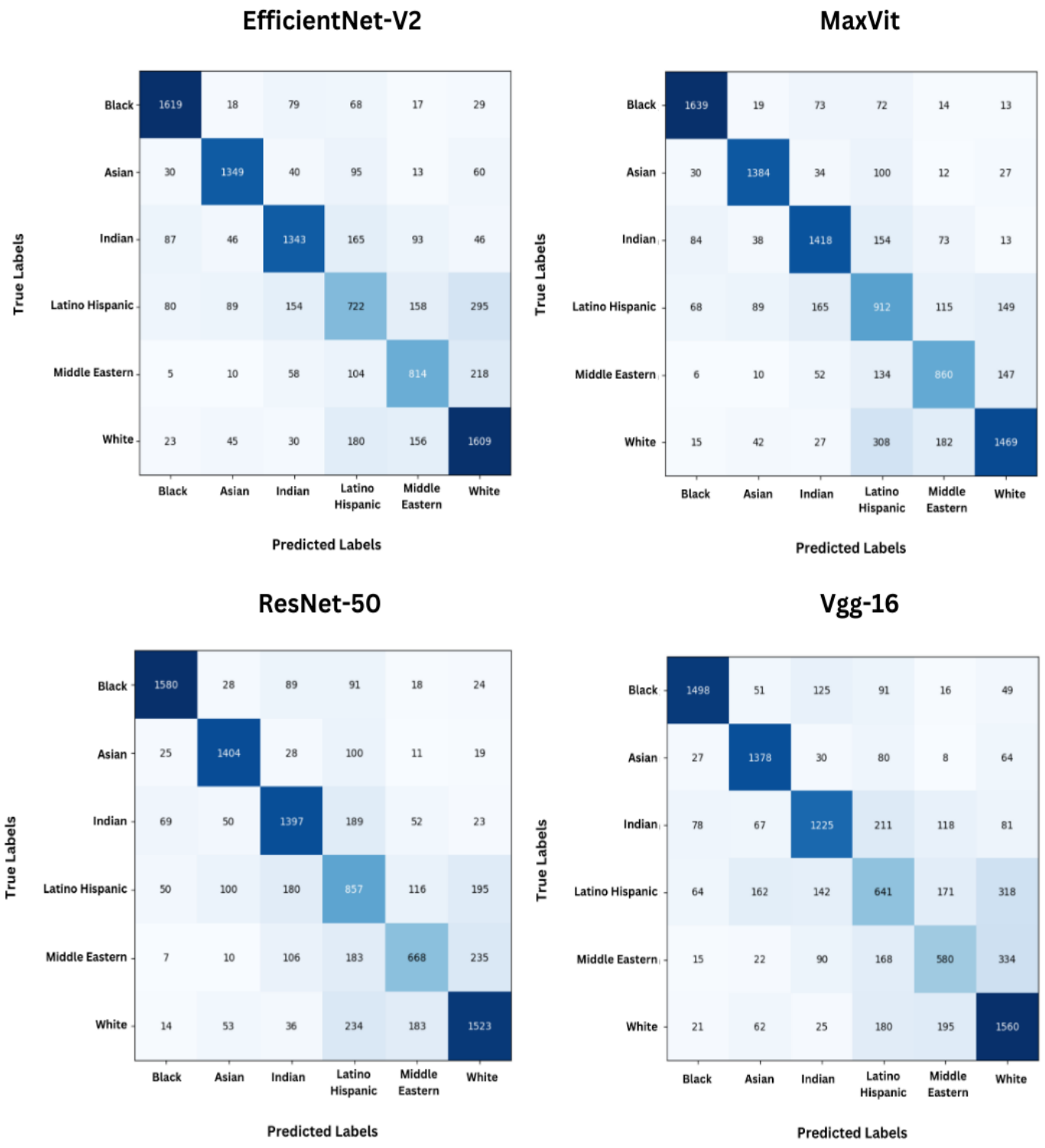

Fig. 9. Confusion matrices.

Based on observation, the model's misclassification for both of Latino Hispanic, and Middle Eastern is noteworthy. This due to the high overlapping diversity between the three ethnic groups of White, Latino Hispanic, and Middle Eastern which are merely considered multiracial groups [43]. Thus, considerable efforts are required for the creation of a representative dataset for such ethnic groups, which certainly could improve the performance of ethnicity recognition models. However, in the conducted experiments the proposed MaxVit model achieves better generalization compared to other models.

V. CONCLUSION

This paper addresses the two common limitations of research in the field of race recognition. First, a large database has been created with six racial categories i.e., Asian, Black, Indian, Latino Hispanic, Middle Eastern, and White. Second, it has proposed the usage of a vision transformer named MaxVit as an ethnicity recognition model using facial images. It achieves a classification accuracy of 77.2% and better generalization than other recent works.

The research area of race and ethnicity recognition is still unsaturated, mainly from the aspect of racial or ethnic groups diversity, as the number of pre-defined racial categories is limited by the available datasets. This could be particularly problematic for individuals who have mixed racial backgrounds, thus multiracial classification is noteworthy as racial facial traits are noticeably overlapped for most of population individuals. Therefore, a considerable effort should be devoted towards such aspects, which certainly contribute to the advancement of the research area.

ACKNOWLEDGMENT

This work was funded by the University of Jeddah, Jeddah, Saudi Arabia, under grant no. (UJ-21-DR-12). The authors, therefore, acknowledge with thanks the University of Jeddah technical and financial support.

REFERENCES

[1] "Definition of Race," Merriam-Webster Dictionary, 2023.

[2] A. Morgan, P. Catherine, P. Heather, "Ethnicity," Oxford Classical Dictionary, Oxford University Press, 2015.

[3] J. Blumenbach, T. Bendyshe, "The anthropological treatises of Johann Friedrich Blumenbach," 1865.

[4] E. Jensen, "Measuring racial and ethnic diversity for the 2020 census," The United States Census Bureau, 2021.

[5] ''Revisions to the standards for the classification of Federal data on race and ethnicity,'' Office of the Federal Register, National Archives and Records Administration, Federal Register 62, no. 210, 58782-58790 1997.

[6] J. Calder, G. Rhodes, M. Johnson, and J. V. Haxby, "Oxford handbook of face perception," 2011.

[7] J.D. Nguyen, H. Duong, "Anatomy, Head and Neck, Face," Treasure Island (FL): StatPearls Publishing, 2023.

[8] A. Greco, G. Percannella, M. Vento, and V. Vigilante, "Benchmarking deep network architectures for ethnicity recognition using a new large face dataset,'' Machine Vision and Applications, 31-67, 2020.

[9] Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-Axis Vision Transformer,'' European Conference on Computer Vision, 2022.

[10] K. Kimmo, and J. Joo, "FairFace: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation," Workshop on Applications of Computer Vision, 2021.

[11] Z. Zhang, Y. Song, and H. Qi, "Age progression/regression by conditional adversarial autoencoder," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp.5810–5818, 2017.

[12] N.A. Al-Humaidan, M. Prince, "A classification of arab ethnicity based on face image using deep learning approach,'' in IEEE Access, vol. 9, pp.50755-50766, 2021.

[13] 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), 2015.

[14] K. Ricanek, T. Tesafaye, "Morph: A longitudinal image database of normal adult age-progression," In 7th International Conference on Automatic Face and Gesture Recognition (FGR06), pp.341–345, 2006.

[15] M. Wang, W. Deng, J. Hu, X. Tao, Y. Huang, ''Racial faces in the wild: reducing racial bias by information maximization adaptation network,'' ICCV, 2019.

[16] M. Wang, Y. Zhang, W. Deng, "Meta balanced network for fair face recognition,'' TPAMI, 2021.

[17] Q. Cao, L. Shen, W. Xie, O.M. Parkhi, and A. Zisserman, ''Vggface2: a dataset for recognising faces across pose and age. In: IEEE International Conference on Automatic Face & Gesture Recognition, pp.67–74, 2018.

[18] M. Merler, N. Ratha, R. S. Feris, and J. R. Smith, ''Diversity in faces,'' arXiv preprint arXiv:1901.10436, 2019.

[19] R. Rothe, R. Timofte, and L. Van Gool, ''Deep expectation of real and apparent age from a single image without facial landmarks,'' International Journal of Computer Vision (IJCV), 2016.

[20] B. Chen, C. Chen, and W. H. Hsu, ''Face recognition and retrieval using cross-age reference coding with cross-age celebrity dataset,'' IEEE Transactions on Multimedia, 17(6):804-815, 2015.

[21] L. Farkas, "Anthropometry of the head and face," Raven Press, 2nd ed., 1994.

[22] X. Lu and A. K. Jain, "Ethnicity identification from face images," Proc. SPIE 5404, Biometric Technology for Human Identification, 2004.

[23] G. Zhang, and Y. Wang, "Multimodal 2D and 3D facial ethnicity classification," 2009 Fifth International Conference on Image and Graphics, 2009.

[24] S. Hosoi, E. Takikawa and M. Kawade, "Ethnicity estimation with facial images," Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004.

[25] N. Narang, T. Bourlai, "Gender and ethnicity classification using deep learning in heterogeneous face recognition," 2016 International Conference on Biometrics (ICB), 2016.

[26] T. Kazimov and S. Mahmudova, "About a method of recognition of race and ethnicity of individuals based on portrait photographs," Intelligent Control and Automation, 5, pp.120-125, 2014.

[27] H. Ding, D. Huang, Y. Wang, and L. Chen, "Facial ethnicity classification based on boosted local texture and shape descriptions," 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 2013.

[28] M. A. Uddin, and S. A. Chowdhury, "An integrated approach to classify gender and ethnicity," 2016 International Conference on Innovations in Science, Engineering and Technology (ICISET), 2016.

[29] M. Obayya, S. S. Alotaibi, S. Dhahb, R. Alabdan, M. Al-Duhayyim, M. A. Hamza, M. Rizwanullah, and A. Motwakel, ''Optimal deep transfer learning based ethnicity recognition on face images,'' Image and Vision Computing, Volume 128, 2022.

[30] G. Sunitha, K. Geetha, S. Neelakandan, A. K. S. Pundir, S. Hemalatha, and V. Kumar, ''Intelligent deep learning based ethnicity recognition and classification using facial images,'' Image and Vision Computing, Volume 121, 2022.

[31] H. Zhao, D. Manandhar and Kim-Hui Yap, "Hybrid supervised deep learning for ethnicity classification using face images," 2018 IEEE International Symposium on Circuits and Systems (ISCAS), 2018.

[32] H. Hu, J. Anil K., W. Fang, S. Shiguang, and C. Xilin, ''Heterogeneous face attribute estimation: A deep multi-task learning approach,'' IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 40, no. 11, pp.2597-2609, 2018.

[33] A. Inzamam, and N. Ul-Islam, "Learned features are better for ethnicity classification," Cybernetics and Information Technologies 17, 2017.

[34] A. Ahmed, K. Yu, W. Xu, Y. Gong, and E. Xing, ''Training hierarchical feed-forward visual recognition models using transfer learning from pseudo-tasks,'' In: Forsyth, D., Torr, P., Zisserman, A. (eds) Computer Vision – ECCV 2008. ECCV, 2008.

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

[36] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.C. Chen, ''Mobilenetv2: Inverted residuals and linear bottlenecks,'' In Proceedings of the IEEE conference on computer vision and pattern recognition, pp.4510-4520, 2018.

[37] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, ''Transformers in vision: A survey'', In ACM Computing Surveys, Association for Computing Machinery (ACM), 2022.

[38] J. Hu, L. Shen, and G. Sun, ''Squeeze-and-excitation networks,'' In Proceedings of the IEEE conference on computer vision and pattern recognition. pp.7132-7141, 2018.

[39] J. Deng, W. Dong, R. Socher, L. -J. Li, K. Li, and Li Fei-Fei, "ImageNet: A large-scale hierarchical image database," 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 2009.

[40] M. D. Zeiler, "ADADELTA: An Adaptive Learning Rate Method," ArXiv, abs/1212.5701, 2012.

[41] P. Omkar M., A. Vedaldi, and A. Zisserman, "Deep face recognition." British Machine Vision Conference, 2015.

[42] M. Tan, and Q. V. Le, '' EfficientNetV2: Smaller models and faster training.'' ArXiv, abs/2104.00298, 2021.

[43] L. Charmaraman, M. Woo, A. Quach, and S. Erkut, ''How have researchers studied multiracial populations? A content and methodological review of 20 years of research,'' Cultural Diversity and Ethnic Minority Psychology, 20(3), 2014.