ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA241024

Device-Specific Facial Descriptors: Winning a Lottery with a SuperNet

Andrey Savchenko[a,b,d,*], **Dmitry Maslov**[b] **and Ilya Makarov**[c,d]

[a]Sber AI Lab, Moscow, Russia
[b]Laboratory of Algorithms and Technologies for Network Analysis, HSE University, Nizhny Novgorod, Russia
[c]Artificial Intelligence Research Institute (AIRI), Moscow, Russia
[d]ISP RAS Research Center for Trusted Artificial Intelligence, Moscow, Russia
ORCID (Andrey Savchenko): https://orcid.org/0000-0001-6196-0564, ORCID (Ilya Makarov): https://orcid.org/0000-0002-3308-8825

Abstract. We address the challenge of devising neural network architectures to extract facial descriptors across diverse mobile and edge devices. Employing neural architecture search, we introduce a novel framework that selects optimal subnetworks from a SuperNet using an evolutionary search. Using a surrogate gradient boosting classifier to avoid direct accuracy estimation of subnetworks on validation sets, our approach swiftly delivers the most efficient and accurate models tailored to specific devices within minutes. Demonstrating versatility through an Android demo app, our framework excels in tasks like face recognition and emotion understanding across various devices, achieving real-time processing and superior accuracy compared to existing mobile models.

1 Introduction

A lot of real-world applications, e.g., human-machine interaction and video surveillance [22, 31], need to solve facial classification tasks, such as facial expression recognition (FER) [6, 12, 32] and face recognition [11, 19]. Due to privacy issues, it is typically required to solve these tasks on-device [4, 10, 25]. However, the landscape of mobile and edge computing is characterized by a rich diversity of devices, each endowed with unique processing capabilities [20, 33, 38]. This heterogeneity poses a significant hurdle in developing a universal neural network architecture for tasks such as facial descriptor extraction [29, 27]. Addressing this challenge head-on, our research studies AutoML techniques to devise tailored neural networks optimized for specific devices [7, 18].

Training a custom descriptor for each device poses significant time constraints. Hence, the central to our study is the concept of a SuperNet [34, 24]. It is a comprehensive Once-for-All architecture encompassing many potential subnetworks with shared weights [7], so it is possible to extract specific subnetworks suitable for a concrete device. A particular procedure to train SuperNet was proposed in [2] based on the Pareto ranking between its subnets. The SuperNet exploits the lottery ticket hypothesis [13] that suggests that large neural networks can be pruned to smaller models with comparable accuracy, emphasizing the need to navigate this pruning process efficiently. Unfortunately, the original Once-for-All network [7] falls short in addressing the requirements of facial processing for generating high-quality descriptors rather than solely optimizing classification accuracy on validation sets.

In this paper, leveraging the concept of SuperNet, we craft a methodology (Fig. 1) that navigates the complexity of device variations to extract accurate facial descriptors. In particular, it harnesses the power of genetic algorithms with surrogate binary classifiers to identify the most promising subnetworks within the SuperNet. This innovative approach bypasses the need for direct accuracy estimation on validation sets, streamlining the model selection process and significantly reducing computational overhead. The result is a neural network architecture optimized for each specific device, tailored to strike the delicate balance between computational efficiency and accuracy of facial classification [14, 15].

The source code of our demo application and several pretrained neural networks are publicly available[1]. The demonstration video for this framework and mobile demo application is available at[2].

2 Methodology

2.1 Proposed Approach

The proposed methodology contains several steps. We begin by training the Once-for-All SuperNet [7, 30, 35] for a specific face classification task. This paper examines two problems, namely, FER and face identification [1]. In the former case, the training part of manually labeled facial photos from the AffectNet dataset [21] was used. Each photo is associated with one of eight classes: Anger, Contempt, Disgust, Fear, Happiness, Neutral, Sadness, and Surprise. The validation part of AffectNet is a balanced set of 4000 images (500 per class). In the latter case, the SuperNet was trained to recognize celebrities from the VGGFace2 dataset [8]. The training and validation sets contain 3,067,564 and 243,722 photos of 9131 celebrities. As a result, we obtained two SuperNets for FER and face recognition, respectively.

Next, extracting the optimal subnetwork suitable to extract facial descriptors on a specific device with a restriction in the inference time $\bar{t}$ is necessary. The architecture of a subnetwork is described

[1] https://github.com/av-savchenko/mobile-face-recognition
[2] https://youtu.be/xE3SHElYzM4

* Corresponding Author. Email: avsavchenko@hse.ru.

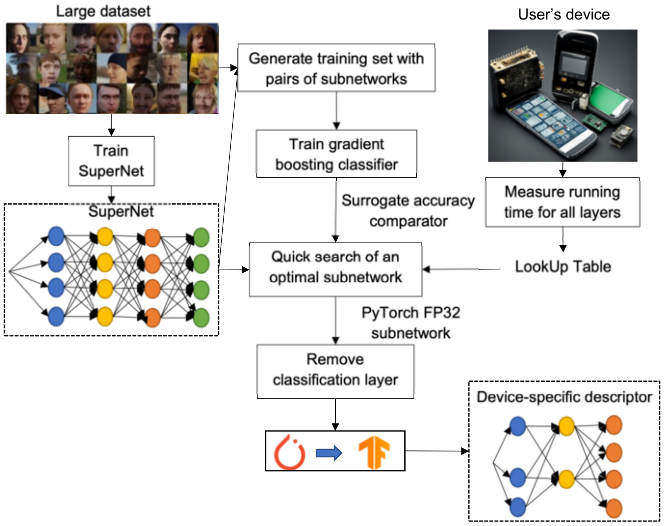

Figure 1. Proposed methodology to obtain device-specific facial descriptors

by the number $d \in \{2,3,4\}$ of layers in each of the five groups of blocks, where each block is a convolutional layer with a kernel size $ks \in \{3,5,7\}$ and a scaling factor $e \in \{3,4,6\}$. Hence, the number of different subnetworks is enormous ($\sim 2 \cdot 10^{19}$), so it is necessary to use evolutionary search to “win the lottery”. Searching for the best subnetwork requires comparing the accuracy of two arbitrary subnetworks. The most obvious way to do it is to estimate the accuracy using a validation set. Unfortunately, this procedure may be very time-consuming, and evolutionary algorithms typically require to compare thousands of subnetworks. As a result, a surrogate classifier is needed.

The authors of the Once-For-All framework [7] trained a multi-layer perceptron to predict the validation accuracy for a given architecture of a subnetwork. Unfortunately, it was noticed that training such a regression model is very difficult as it usually overestimates the predicted accuracy. In this paper, we propose to compare the accuracy of two subnetworks with a special binary gradient-boosting classifier. We generate a diverse training set of 16000 random subnetworks, estimate their accuracy on a validation set, and train the surrogate binary classifier (LightGBM) to determine the relative accuracy of two given subnetworks.

To facilitate hardware-specific optimization, we measure the running time of each layer of the Once-for-All network on a concrete device. Leveraging the obtained Look-Up Tables (LUTs) alongside the trained gradient boosting classifier and maximal inference time $\bar{t}$, we implement a genetic algorithm to select the subnetwork with maximal expected accuracy while meeting latency requirements. By utilizing a QuickSelect partition algorithm, this algorithm ensures linear complexity dependent on the number of iterations and population size, expediting the search process significantly.

Finally, the last classification layer of the selected subnetwork is removed. The resulting model is implemented in PyTorch with FP32 weights. PyTorchMobile is now 1.5-2 times slower than TensorFlow Lite. Unfortunately, automatic conversion from PyTorch to ONNX leads to rather slow models. Hence, we implemented custom scripts to create a subnetwork from scratch in TensorFlow and then copy the weights from PyTorch to TensorFlow format. The latter is automatically converted to TensorFlowLite for deployment on mobile devices.

2.2 Our Framework

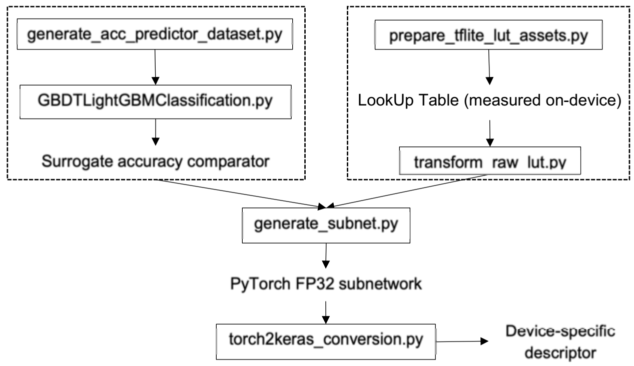

Figure 2. Our framework

The proposed approach was implemented in a special Python framework (Fig. 2) that lets the user 1) generate a dataset for accuracy prediction and train the surrogate binary classifier; 2) prepare LUT for each OFA's layer; and 3) generate subnetwork in PyTorch format given latency constraint and convert it to TensorFlowLite model. In addition, a special face_rec_model_tester Jupyter notebook is available to test the quality of subnetworks. Our demo makes SuperNets for FER and face identification publicly available. In our experiments, we used the Raspberry Pi 4 mini-computer and two mobile devices with Android: Xiaomi Mi 10T with Qualcomm Snapdragon 865 and Xiaomi Mi 10 Lite with Snapdragon 765g. We extracted two subnetworks by setting the maximal inference time relative to the inference time $\bar{t}_{ENet}$ of the EfficientNet-B0 (TFLite) model. As a result, we obtained subnetwork 1 and 2 for $0.6\bar{t}_{ENet}$ and $0.4\bar{t}_{ENet}$, respectively.

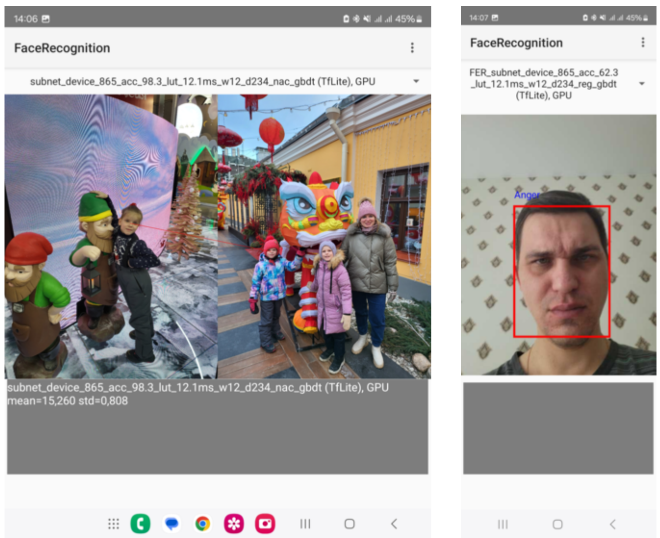

Figure 3. Sample UI of our demo application

Finally, we developed a special demo application for Android devices (Fig. 3). The source code is made publicly available in our GitHub repository. It supports the following functionality. First, we support facial matching on two photos from a mobile device gallery. Here, all faces are detected on both photos and facial descriptors are extracted and compared mutually. The red line is drawn between corresponding faces if the cosine similarity between descriptors is higher than a predefined threshold. Secondly, it is possible to capture the frontal camera, detect facial region using MTCNN [37], and recognize facial expressions with one of the selected subnetworks. Thirdly, we support measuring the average inference time of several subnetworks and existing models for facial feature extraction.

3 Experimental results

3.1 Facial Expression Recognition

In the first experiment, the validation set of the AffectNet dataset is used. We compare our subnetworks with baseline AlexNet [21], SL+SSL inpanting (EfficientNet-B0) [23], ViT-base + MAE [17], and EmotiEffNet-B0/B2 [26]. Table 1 contains the validation accuracy α for 8 classes and average CPU running times $\bar{t}_R, \bar{t}_{865}, \bar{t}_{765}$ for Raspberry Pi4, Xiaomi Mi 10T and Xiaomi Mi 10 Lite, respectively. As one can notice, our lightweight models found an ideal balance between speed and accuracy. It is worth mentioning that despite low inference time of the baseline AlexNet, it has 3-5 times greater number of weights than our models, which may be very important for edge and mobile devices [33].

Table 1. Facial expression recognition accuracy α (%) on AffectNet and mean inference time per one face $\bar{t}_R, \bar{t}_{865}, \bar{t}_{765}$ (ms) for Raspberry Pi 4 and mobile devices with Snapdragon 865, Snapdragon 765, respectively

Model	α, %	$\bar{t}_R$, ms	$\bar{t}_{865}$, ms	$\bar{t}_{765}$, ms
AlexNet (baseline)	58.0	62.51	8.33	17.01
EmotiEffNet-B0	61.32	183.15	47.15	122.76
SSL inpanting	61.72	183.61	47.32	123.07
ViT-base + MAE	62.42	1084.93	487.21	952.50
EmotiEffNet-B2	63.03	358.32	149.87	381.12
Our subnetwork 1	62.05	108.14	11.93	34.02
Our subnetwork 2	61.28	73.26	8.90	22.78

3.2 Face Recognition

In the second experiment, we used our models for face identification on the LFW (Labeled Faces in the Wild) dataset [16]. We used the conventional protocol [3], which selects 596 subjects with at least two photos in the LFW and at least one video in the YouTube Faces database. One facial photo of each subject is copied into the training set; the validation set contains all other photos. The average accuracy of the 1-NN classifier computed using five times randomly repeated cross-validation is presented in Table 2. Here, we compare two techniques to crop the facial region after face detection [28]:

1. Alignment with similarity transform and conversion to 224x224, in which background is available.
2. Simple crop of detected faces without any margins.

Our models are compared with traditional InsightFace (IResNet-50) [11], VGGFace2 (SENet-50) [8], FaceNet (InceptionResNet) [29], PocketNetM-256 [5], MobileFaceNet [9] and EfficientNet-B0/B2 [26]. As one can notice, our methodology lets us obtain the fastest models, which show high accuracy for various facial preprocessing techniques. Such reliability is an important factor, as specific backgrounds in aligned faces may significantly influence the quality of facial descriptors.

In the final experiment (Table 3), we compare our surrogate binary classifier with an accuracy predictor trained as described by the authors of the Once-for-All approach [7]. The proposed methodology yields enhancements in face recognition accuracy, with improvements of up to 0.9%, particularly under less stringent time constraints. Despite incorporating a more intricate genetic algorithm based on the quick sort and employing a binary gradient boosting classifier instead of a simple multi-layered feed-forward neural network, our search time $\bar{t}_S$ only marginally increases by 30 seconds. Nonetheless, our search duration remains under five minutes, even with this adjustment.

Table 2. Face identification accuracy α_1, α_2 (%) on LFW for aligned and cropped faces, respectively, mean inference time per one face $\bar{t}_{865}$ (ms) for a mobile device with Snapdragon 865 and size of the model M (Mb).

Model	α_1, %	α_2, %	$\bar{t}_{865}$, ms	M, Mb
InsightFace	99.23	82.34	203.75	166
VGGFace2	97.21	96.61	123.07	167
FaceNet	96.12	96.57	110.63	107
PocketNetM-256	99.70	76.12	407.86	7
MobileFaceNet	97.42	44.23	13.28	6
EfficientNet-B0	94.07	94.70	47.07	16
EfficientNet-B2	95.00	91.53	148.70	30
Our SuperNet	98.97	99.12	34.02	34
Our subnetwork 1	98.13	98.71	11.89	18
Our subnetwork 2	96.89	97.34	8.74	13

Table 3. Comparison of the proposed approach with OFA: face identification accuracy α for cropped LFW faces and time of search $\bar{t}_S$ (minutes).

		α, %		$\bar{t}_S$, min.	
Device	Constraint	OFA	Ours	OFA	Ours
Snapdragon 865	$t \leq 0.6t_{ENet}$	97.95	98.71	0.05	0.57
	$t \leq 0.4t_{ENet}$	96.89	97.34	0.20	1.02
Snapdragon 765	$t \leq 0.6t_{ENet}$	97.94	98.84	0.72	1.35
	$t \leq 0.4t_{ENet}$	96.89	96.91	4.14	4.89
Raspberri Pi4	$t \leq 0.6t_{ENet}$	98.51	98.78	0.22	0.65
	$t \leq 0.4t_{ENet}$	97.29	97.30	0.19	0.87

4 Conclusion

In this paper, we have proposed the framework (Fig. 1) to efficiently generate optimized neural networks for facial feature extraction tailored to specific hardware and latency constraints. To demonstrate the efficacy and versatility of our approach, we have developed an Android demo application (Fig. 3). The source code of our demo application and trained FER models are publicly available, while the code of each step of our methodology will be made available after peer review. It was experimentally shown that our models exhibit superior performance across a spectrum of devices, ranging from smartphones to Raspberry Pi, from face recognition to facial expression recognition. By achieving real-time processing and outperforming existing lightweight networks in accuracy, our research underscores the transformative potential of device-specific neural networks in shaping the future of mobile and edge computing applications.

The primary limitation of the proposed approach lies in its inability to ensure the high quality of extracted facial descriptors. Moreover, our evaluation metrics primarily focus on accuracy and inference time, potentially overlooking other important aspects such as robustness, generalization, or interpretability of the models [36]. The final drawback of our subnetworks is a high number of parameters (Table 2). Hence, in the future, it is necessary to choose more space-efficient layers and incorporate such techniques as ArcFace into SuperNet training.

Acknowledgements

The article was prepared within the framework of the Basic Research Program at the National Research University Higher School of Economics (HSE).

References

[1] N. S. Belova and A. V. Savchenko. Unconstrained face identification using maximum likelihood of distances between deep off-the-shelf features. *Expert Systems with Applications*, 108:170–182, 2018.

[2] H. Benmeziane, K. E. Maghraoui, H. Ouarnoughi, and S. Niar. Pareto rank-preserving supernetwork for hardware-aware neural architecture search. In *Proceedings of European Conference on Artificial Intelligence (ECAI)*, 2023.

[3] L. Best-Rowden, H. Han, C. Otto, B. F. Klare, and A. K. Jain. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Transactions on Information Forensics and Security*, 9(12):2144–2157, 2014.

[4] A. Boragule, K. C. Yow, and M. Jeon. On-device face authentication system for ATMs and privacy preservation. In *Proceedings of International Conference on Consumer Electronics (ICCE)*, pages 1–4. IEEE, 2023.

[5] F. Boutros, P. Siebke, M. Klemt, N. Damer, F. Kirchbuchner, and A. Kuijper. PocketNet: Extreme lightweight face recognition network using neural architecture search and multistep knowledge distillation. *IEEE Access*, 10:46823–46833, 2022.

[6] S. Buechel and U. Hahn. Emotion analysis as a regression problem–dimensional models and their implications on emotion representation and metrical evaluation. In *Proceedings of European Conference on Artificial Intelligence (ECAI)*, pages 1114–1122. IOS Press, 2016.

[7] H. Cai, C. Gan, T. Wang, Z. Zhang, and S. Han. Once-for-All: Train one network and specialize it for efficient deployment. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–15, 2020.

[8] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman. VggFace2: A dataset for recognising faces across pose and age. In *Proceedings of the 13th IEEE international conference on automatic face & gesture recognition (FG)*, pages 67–74. IEEE, 2018.

[9] S. Chen, Y. Liu, X. Gao, and Z. Han. MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices. In *proceedings of the 13th Chinese Conference on Biometric Recognition (CCBR)*, pages 428–438. Springer, 2018.

[10] P. Demochkina and A. V. Savchenko. MobileEmotiFace: Efficient facial image representations in video-based emotion recognition on mobile devices. In *Proceedings of ICPR International Workshops and Challenges ob Pattern Recognition, Part V*, pages 266–274. Springer, 2021.

[11] J. Deng, J. Guo, T. Liu, M. Gong, and S. Zafeiriou. Sub-center ArcFace: Boosting face recognition by large-scale noisy web faces. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, pages 741–757. Springer, 2020.

[12] Z. Du, X. Jiang, P. Wang, Q. Zhou, X. Wu, J. Zhou, and Y. Wang. LION: label disambiguation for semi-supervised facial expression recognition with progressive negative learning. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 699–707, 2023.

[13] J. Frankle and M. Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2019.

[14] A. Karpov and I. Makarov. Exploring efficiency of vision transformers for self-supervised monocular depth estimation. In *Proceedings of International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 711–719. IEEE, 2022.

[15] A. Kharchevnikova and A. Savchenko. Neural networks in video-based age and gender recognition on mobile platforms. *Optical Memory and Neural Networks*, 27:246–259, 2018.

[16] E. Learned-Miller, G. B. Huang, A. RoyChowdhury, H. Li, and G. Hua. Labeled faces in the wild: A survey. In *Proceedings of the International Conference on Advances in Face Detection and Facial Image Analysis*, pages 189–248. Springer, 2016.

[17] J. Li, J. Nie, D. Guo, R. Hong, and M. Wang. Emotion separation and recognition from a facial expression by generating the poker face with vision transformers. *arXiv preprint arXiv:2207.11081*, 2023.

[18] C.-H. Liu, Y.-S. Han, Y.-Y. Sung, Y. Lee, H.-Y. Chiang, and K.-C. Wu. FOX-NAS: fast, on-device and explainable neural architecture search. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 789–797, 2021.

[19] D. Liu, N. Wang, C. Peng, J. Li, and X. Gao. Deep attribute guided representation for heterogeneous face recognition. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 835–841, 2018.

[20] Z. Liu, G. Lan, J. Stojkovic, Y. Zhang, C. Joe-Wong, and M. Gorlatova. Collabar: Edge-assisted collaborative image recognition for mobile augmented reality. In *Proceedings of the 19th International Conference on Information Processing in Sensor Networks (IPSN)*, pages 301–312. IEEE, 2020.

[21] A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, 2017.

[22] S. Nayak, B. Nagesh, A. Routray, and M. Sarma. A human–computer interaction framework for emotion recognition through time-series thermal video sequences. *Computers & Electrical Engineering*, 93:107280, 2021.

[23] M. Pourmirzaei, G. A. Montazer, and F. Esmaili. Using self-supervised auxiliary tasks to improve fine-grained facial representation. *arXiv preprint arXiv:2105.06421*, 2021.

[24] S. Sarti, E. Lomurno, A. Falanti, and M. Matteucci. Enhancing Once-For-All: A study on parallel blocks, skip connections and early exits. In *Proceedings of International Joint Conference on Neural Networks (IJCNN)*, pages 1–9. IEEE, 2023.

[25] A. Savchenko. Facial expression recognition with adaptive frame rate based on multiple testing correction. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pages 30119–30129. PMLR, 2023.

[26] A. V. Savchenko. EmotiEffNets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5716–5724, 2023.

[27] A. V. Savchenko. AutoFace: How to obtain mobile neural network-based facial feature extractor in less than 10 minutes? *IEEE Access*, 12:25106–25118, 2024.

[28] A. V. Savchenko, L. V. Savchenko, and I. Makarov. Fast search of face recognition model for a mobile device based on neural architecture comparator. *IEEE Access*, 11:65977–65990, 2023.

[29] F. Schroff, D. Kalenichenko, and J. Philbin. FaceNet: A unified embedding for face recognition and clustering. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823. IEEE, 2015.

[30] J. Shipard, A. Wiliem, and C. Fookes. Does interference exist when training a Once-For-All network? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3619–3628, 2022.

[31] A. D. Sokolova, A. S. Kharchevnikova, and A. V. Savchenko. Organizing multimedia data in video surveillance systems based on face verification with convolutional neural networks. In *Proceedings of the 6th International Conference on Analysis of Images, Social Networks and Texts (AIST)*, pages 223–230. Springer, 2018.

[32] E. Veltmeijer, C. Gerritsen, and K. Hindriks. Automatic recognition of emotional subgroups in images. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1363–1370, 2022.

[33] P. Warden and D. Situnayake. *TinyML: Machine learning with TensorFlow Lite on Arduino and ultra-low-power microcontrollers*. O'Reilly Media, 2019.

[34] H. You, B. Li, Z. Sun, X. Ouyang, and Y. Lin. SuperTickets: Drawing task-agnostic lottery tickets from supernets via jointly architecture searching and parameter pruning. In *Proceedings of European Conference on Computer Vision (ECCV), Part XI*, pages 674–690. Springer, 2022.

[35] K. Yu, R. Ranftl, and M. Salzmann. How to train your Super-Net: An analysis of training heuristics in weight-sharing nas. *arXiv preprint arXiv:2003.04276*, 2020.

[36] J. Zhang and H. Yu. Improving the facial expression recognition and its interpretability via generating expression pattern-map. *Pattern Recognition*, 129:108737, 2022.

[37] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[38] Y. Zhang, F. Wang, W. Sun, J. Su, P. Liu, Y. Li, X. Feng, and Z. Zou. Matting moments: a unified data-driven matting engine for mobile AIGC in photo gallery. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI)*, pages 7183–7186, 2023.