Face Attribute Detection with MobileNetV2 and NasNet-Mobile

Conference Paper · September 2019

DOI: 10.1109/ISPA.2019.8868585

6 authors, including:

Frerk Saxen
Otto-von-Guericke University Magdeburg
37 PUBLICATIONS **497** CITATIONS
SEE PROFILE

Philipp Werner
Otto-von-Guericke University Magdeburg
71 PUBLICATIONS **2,501** CITATIONS
SEE PROFILE

Sebastian Handrich
Otto-von-Guericke University Magdeburg
29 PUBLICATIONS **267** CITATIONS
SEE PROFILE

Ehsan Othman
Otto-von-Guericke University Magdeburg
12 PUBLICATIONS **232** CITATIONS
SEE PROFILE

Face Attribute Detection with MobileNetV2 and NasNet-Mobile

Frerk Saxen, Philipp Werner, Sebastian Handrich, Ehsan Othman, Laslo Dinges, Ayoub Al-Hamadi
Faculty of Electrical Engineering and Information Technology, Neuro-Information Technology
Otto von Guericke University
Magdeburg, Germany
Frerk.Saxen@ovgu.de

***Abstract*—In this paper, we propose two simple yet effective methods to estimate facial attributes in unconstrained images. We use a straight forward and fast face alignment technique for preprocessing and estimate the face attributes using MobileNetV2 and Nasnet-Mobile, two lightweight CNN (Convolutional Neural Network) architectures. Both architectures perform similarly well in terms of accuracy and speed. A comparison with state-of-the-art methods with respect to processing time and accuracy shows that our proposed approach perform faster than the best state-of-the-art model and better than the fastest state-of-the-art model. Moreover, our approach is easy to use and capable of being deployed on mobile devices.**

***Index Terms*—Mobile face attribute detection, MobileNetV2, Nasnet-Mobile**

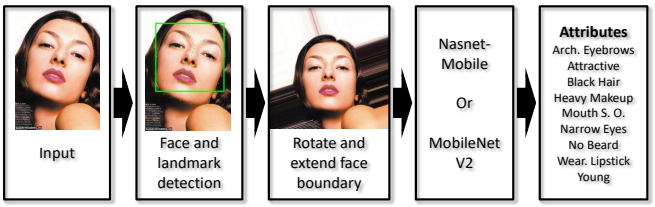

Fig. 1: Proposed attribute estimation pipeline. The face and landmark detection as well as the Nasnet-Mobile and MobileNetV2 implementation are available online at [7], [8], [9] and [10], respectively. The trained models will be provided on request.

I. Introduction

Estimating human face attributes is important for several applications (e.g. face retrieval, social media or video surveillance [1]). The estimation however is difficult due to vast changes in appearance and shape of different attributes, out of plane head rotations and difficult lighting conditions in unconstrained settings. Many applications, however, require accurate and fast solutions that perform on resource-constrained systems, such as mobile devices. E.g. it may not be possible to upload images to the cloud for recognition due to privacy reasons or a lack of Internet connection.

In this work we discuss previous works on facial attribute detection and evaluate two lightweight CNN architectures with respect to performance and speed using a straight forward methodology that can be implemented on mobile devices.

A. Related Work

During the last decade, multiple approaches have been proposed for face attribute estimation. Similar to other domains, deep learning approaches excelled traditional methods with the emergence of huge datasets like CelebA [2].

LNets+ANet: Liu et al. [2] propose a combined face and attribute detection framework (similar to R-CNN [3]). Thus it does not require any preprocessing like face and landmark detection. However, the pipeline takes a considerable amount of training time but the results made a remarkable improvement over the state-of-the-art. With the introduction of Faster R-CNN [4] and YOLOv3 [5] (for object detection) many shortcomings have been eliminated. Nevertheless, since Liu et al. [2] nobody tried to adopt these models for face attribute detection again.

MCNN-AUX: Hand and Chellappa [6] propose a multi-task CNN with an auxiliary network. They suggest a straight forward CNN architecture and manually group the attributes to train the multi-task CNN. The architecture has about 64 million parameters but only 3 convolution layer and 2 fully connected layer. The training was done within 2.5 hours only. Thus, we suspect a very fast inference time on modern GPUs.

Mid-Level: Zhong et al. [11] propose to use mid-level representations of the pre-trained FaceNet NN.1 [12] architecture. They apply multi-scale spatial pooling on intermediate layers and classify attributes at each level using a Support Vector Machine. For each attribute the best performing layer is chosen. The authors report that "features from the intermediate layers demonstrate an obvious advantage [...] for attributes describing motions of the mouth area where the gap is almost 20%". Although this approach is interesting, it is not trained end-to-end.

AFFACT: Günther et al. [13] uses a ResNet-50 architecture. To facilitate an alignment free attribute detection they perform heavy augmentation during inference by applying 162 different transformations to the detected bounding boxes. Thus, the input tensor of their network has 162 channels. Although they report superior classification results, it should be noted that performing these transformations beforehand is quite costly.

DMTL: Wang et al. [14] and DMTL+: Han et al. [1] propose, similar to Hand and Chellappa [6], a deep multi-task learning approach. They use a slightly modified AlexNet [15]

Saxen et al., "Face Attribute Detection with MobileNetV2 and NasNet-Mobile," International Symposium on Image and Signal Processing and Analysis (ISPA), 2019, DOI: 10.1109/ISPA.2019.8868585.

and model both attribute correlation and attribute heterogeneity in a single network. Instead of the 6 subnetworks by Hand and Chellappa [6] they use 8 subnetworks, one holistic subnetwork and seven local nominal subnetworks.

SPLITFACE: Mahbub et al. [16] propose a CNN architecture that is designed for face attribute detection of partial occluded faces. They create severe occlusions synthetically on CelebA and train a CNN on local patches at facial key points of the image to robustly detect face attributes despite heavy occlusions. However, since we do not address occlusion explicitly we did not include SPLITFACE in our evaluation.

B. Dataset

The availability of comprehensive and well-designed databases is crucial for any classification problem. In this work, we use the CelebA dataset [2]. CelebA is a large-scale dataset of celebrity images with large pose variations and background clutter. In total, there are 202,599 face images of 10,777 identities. The dataset is split into training, validation and test sets, with 162,770, 19,867, and 19,962 images, respectively. For each image, 40 face attributes were annotated. These attributes range from general attributes like sex, age and demographic information to specific and individual characteristics, like e.g. face shape, lip and nose size. All these labels are binary labels, i.e. the corresponding facial attribute is either present or not. Table I shows the performance of different models, including a trivial classifier, that always votes for the majority class. Thus, the trivial classifier can show the label distribution of the test set (if we know the majority class). E.g., 50% of the faces are smiling, and only 2.1% are bold. For the majority of the classes, the distribution is imbalanced. Not all given labels are correct, though. For some attributes, a clear and objective decision might be difficult. However, Rolnick et al. discovered that Deep Learning techniques are highly robust against such label noise [17].

II. Method

In this work, we evaluate the performance of two new architectures for face attribute estimation: Nasnet-Mobile [18] (see Sec. II-B) and MobileNetV2 [19] (see Sec. II-C). Compared to most competitors our training procedure is straight forward (see Fig. 1): We do not modify the network architecture and we do not perform a sophisticated alignment. We discuss the different approaches and dataset specifics. Our proposed methods perform faster than the best state-of-the-art model and better than the fastest state-of-the-art model (see details in Sec. III). Sec. II-A explains the training procedure in detail that is the same for both architectures. In Sec. II-B and Sec. II-C the training details are given for Nasnet-Mobile and MobileNetV2, respectively.

A. Training

As the first step in our recognition pipeline we detect the face with a multi-scale ResNet model [7] and estimate facial landmarks using an ensemble of regression trees [8]. We rotate each image to align the eyes horizontally and crop the face (centered between the eyes) with a square bounding box of 2 times the width of the face detection. We do that to capture details like hair and necklace. Finally, we rescale the crop to a resolution of 256×256 pixels.

To augment the training data we crop the image randomly within a range of 95% to 100% of the respective axes, independently for the image width and height. We randomly flip the image horizontally and randomly change the saturation and value within the HSV color space. After augmentation the image is rescaled to 224×224 and converted to RGB to meet the desired input shape of the CNNs. Also the pixel values are scaled to the required range between -1 and 1. We monitored the validation set accuracy during training and applied the test set only a single time when the training finished.

B. Nasnet-Mobile

Nasnet is a scalable CNN architecture (constructed by neural architecture search) that consists of basic building blocks (cells) that are optimized using reinforcement learning [18]. A cell consists of only a few operations (several separable convolutions and pooling) and is repeated multiple times according to the required capacity of the network. The mobile version (Nasnet-Mobile) consists of 12 cells with 5.3 million parameters and 564 million multiply-accumulates (MACs).

We perform transfer learning with the pre-trained model (pre-trained on ImageNet [20]) from [9] with the suggested training setup (*dropout* $= 0.5$, *weight decay* = $4e-5$, *batch norm decay* $= 0.9997$, *batch norm epsilon* $= 1e-3$). We start with the learning rate of 0.05 (*batch size* $= 64$) and automatically reduce the learning rate until $5e-6$. Instead of the cosine learning decay [21] used by [9] we use an automatic learning rate scheduler [22] that estimates the slope of the loss and reduce the learning rate (by a factor of 0.1) when the loss has not been improved over the last $5k$ training steps. This reduced the training time significantly without sacrificing performance. We trained Nasnet-Mobile for 32 hours on a single NVidia 1080 GTX.

C. MobileNetV2

MobileNetV2 is a CNN architecture for mobile devices proposed by Sandler et al. [19]. Its first version was also designed for face attribute detection but trained and evaluated on Googles inhouse dataset [23]. They introduce inverted residuals and linear bottlenecks and achieve state-of-the-art results balancing inference time and performance for common benchmarks like ImageNet [20], COCO [24], and VOC [25]. Our version of MobileNetV2 has 3.47 million parameters and 300 million MACs.

We used the pre-trained model *mobilenet_v2_1.0_224* from [10] with the suggested training setup (*depth multiplier* $= 1.0$). Just like the Nasnet-Mobile training we start with the learning rate of 0.05 (*batch size* $= 64$) and automatically reduce the learning rate until $5e-6$ (each step with a factor of 0.1). We also tried the pre-trained model with *depth multiplier* $= 1.4$ (*mobilenet_v2_1.4_224*) but it was overfitting. Training MobileNetV2 took 13 hours on our NVidia 1080 GTX.

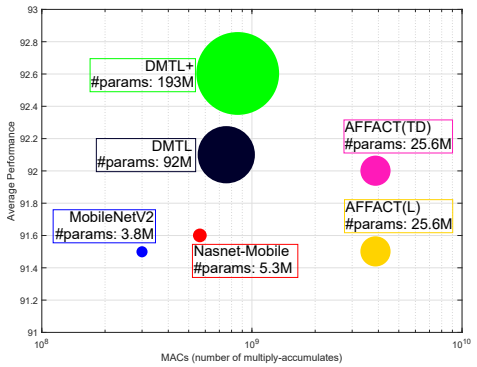

Fig. 2: Number of multiply-accumulates (MACs) needed to compute an inference on a single image vs. average test set accuracy in a log-linear scale. The area of each model is proportional to the number of parameters. Note that we estimated the number of MACs and parameters for DMTL+ [1], DMTL [14], and AFFACT [13] based on their reported changes to well known architectures.

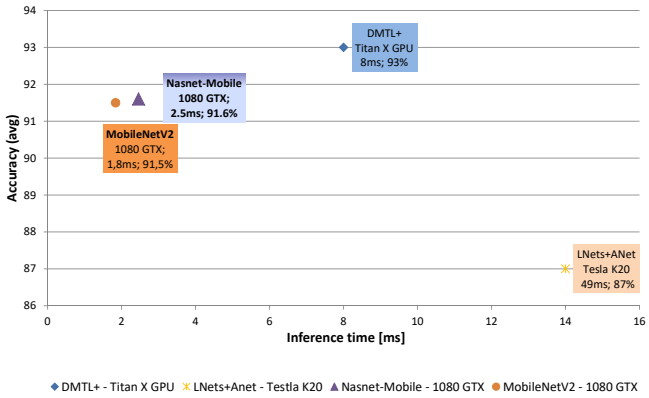

Fig. 3: Mean accuracy vs. inference time for the proposed and other state-of-the-art approaches that report inference time. The results for DMTL+ and LNets+ANet are taken from Han et al. [1].

III. Experiments and Results

Table I shows a comparison of a trivial classifier, 5 state-of-the-art methods and our proposed methods: MobileNetV2 and Nasnet-Mobile. The results of the cited methods are obtained from the respected publications. We report the test set accuracies for the CelebA dataset. The trivial classifier always votes for the majority class (obtained from the training set and applied to the test set). We sorted the methods by their average accuracy.

Although LNets+ANet [2] and Mid-Level [11] are outperformed by more recent methods, both provide interesting solutions that might provide some insights for further research. LNets+ANet showed that a combined face and attribute detection (inspired by R-CNN) can provide excellent results. Future research might adopt recent changes to the object detection methods (e.g. YOLOv3 [5]). Mid-Level [11] uses feature representations from intermediate layers and obtains remarkable results.

MCNN-AUX [6], DMTL [14], and DMTL+ [1] use a multi-task learning approach, where the network is split into several subnetworks. This idea seems to provide promising results. However, from our perspective, it is not clear why splitting the network into several subnetworks improves the performance. A fully connected layer can learn any linear mapping that two separate parallel fully connected layer can learn. Future research might want to investigate this.

While AFFACT (L) [13] uses images aligned with automatically detected landmarks, AFFACT (TD) performs 162 transformations of detected bounding boxes without performing any alignment. This might improve the performance slightly (approx. 0.5%) but comes with a huge computational burden.

Our methods perform similar to MCNN-AUX, AFFACT(L), and DMTL, but are outperformed by DMTL+ and AFFACT(TD). Nevertheless, Fig. 3 shows that our methods perform faster than DMTL+ [1] and LNets+ANet [2] regarding the average inference time for a single image without alignment. The results for DMTL+ and LNets+ANet are from [1]. Note that each approach is evaluated on a different device, thus this comparison is biased. However, other papers did not provide the number of multiply-accumulates (MACs), which would allow for a better comparison. Thus, we estimate the number of MACs and parameters based on their reported changes to well known architectures. DMTL+, and DMTL use a modified AlexNet, and AFFACT uses ResNet-50 (with one additional layer). In Fig. 2 we report the number of MACs vs. average test set accuracy in a log-linear-scale. The area of each circle is proportional to the number of parameters of each model and also proportional to the required memory. This is particularly interesting for mobile applications with limited resources, because storing huge amounts of parameters in memory might not be possible. Compared to DMTL+, our MobileNetV2 model needs 2.9 times less MACs, requires 56 times less memory, and performs 1.1% worse in terms of accuracy. Although Nasnet-Mobile und MobileNetV2 perform very similar, Nasnet-Mobile requires 1.5M more parameters and is about 40% slower. Even though we tried several regularization methods, Nasnet-Mobile was not able to utilize its higher capacity. Usually, models with higher capacity just need more regularization to improve performance. Nevertheless, the test devices weren't too different and even if all methods would have been tested on the same device, we do not believe that their order would change.

IV. Conclusion and Future Work

We addressed the problem of estimating facial attributes from RGB images for mobile devices. Face attribute estimation is important for human machine interaction (HCI) systems by providing relevant context information like age, sex and ethnicity of the interacting user. We evaluated the face attribute detection performance on two mobile architectures: Nasnet-Mobile and MobileNetV2. Our experimental evalua-

	5 o Clock Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	High Cheekbones	Male
Trivial	90.0	71.6	49.6	79.7	97.9	84.4	67.3	78.8	72.8	86.7	94.9	82.0	87.0	94.7	95.4	93.5	95.4	96.8	59.5	51.8	61.4
LNets+ANet [2]	91	79	81	79	98	95	68	78	88	95	84	80	90	91	92	99	95	97	90	87	98
Mid-Level [11]	93.3	82.5	80.8	82.2	97.8	95.6	69.9	82.6	86.0	94.9	96.2	84.2	91.9	94.9	96.2	99.5	97.1	97.8	90.1	86.1	98.1
MCNN-AUX [6]	94.5	83.4	83.1	84.9	98.9	96.0	71.5	84.5	89.8	96.0	96.2	89.2	92.8	95.7	96.3	99.6	97.2	98.2	91.5	87.6	98.2
AFFACT (L) [13]	94.8	83.9	82.8	85.2	99.1	96.1	72.5	84.4	90.5	96.2	96.0	88.5	92.3	95.7	96.4	99.6	97.5	98.3	92.0	87.6	98.2
AFFACT (TD) [13]	94.4	85.5	81.4	84.2	99.0	95.5	84.0	83.0	91.6	95.7	96.1	85.7	92.8	95.7	96.8	99.5	96.7	98.1	91.7	88.3	98.7
DMTL [14]	94	86	83	85	99	96	79	85	91	96	96	88	92	96	97	99	97	98	92	88	98
DMTL+ [1]	95	86	85	85	99	99	96	85	91	96	96	88	92	96	97	99	99	98	92	88	98
MobileNetV2	94.9	84.2	82.7	85.6	99.0	96.2	72.2	84.6	89.9	96.0	96.3	88.8	92.8	95.8	96.5	99.6	97.5	98.3	91.8	87.7	98.4
Nasnet-Mobile	94.9	84.1	83.2	85.4	99.1	96.2	72.3	85.0	90.2	96.1	96.4	89.4	92.9	95.9	96.5	99.7	97.6	98.2	92.0	87.8	98.4

	Mouth Slightly Open	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young		Average
Trivial	50.5	96.1	85.1	85.4	70.4	95.8	71.4	91.5	92.8	95.4	50.0	79.0	63.6	79.3	95.8	47.8	86.2	93.0	75.7		79.9
LNets+ANet [2]	92	95	81	95	66	91	72	89	90	96	92	73	80	82	99	93	71	93	87		87
Mid-Level [11]	92.6	96.6	86.9	95.4	70.6	96.7	76.2	92.1	94.3	97.4	92.1	80.0	77.3	86.7	98.8	92.3	85.8	94.4	87.5		89.8
MCNN-AUX [6]	93.7	96.9	87.2	96.0	75.8	97.0	77.5	93.8	95.2	97.8	92.7	83.6	83.9	90.4	99.0	94.1	86.6	96.5	88.5		91.3
AFFACT (L) [13]	93.8	97.0	87.6	96.2	76.7	97.1	77.1	93.7	95.2	97.8	92.8	85.0	85.7	91.0	99.1	93.7	88.3	96.9	88.8		91.5
AFFACT (TD) [13]	93.9	96.4	93.8	96.0	76.8	96.8	77.6	94.9	95.2	97.3	92.9	85.5	87.9	92.0	98.9	92.7	90.3	96.8	88.7		92.0
DMTL [14]	94	97	90	96	78	97	78	94	96	98	93	85	87	91	99	93	89	97	90		92.1
DMTL+ [1]	94	97	90	97	78	97	78	94	96	98	94	85	87	91	99	93	89	97	90		92.6
MobileNetV2	94.1	97.1	87.8	96.5	76.0	96.8	77.4	93.6	95.1	97.9	93.1	84.6	85.0	90.8	99.1	93.9	87.4	96.8	88.4		91.5
Nasnet-Mobile	94.1	97.1	87.6	96.4	76.4	97.0	77.8	94.0	95.2	98.0	93.1	85.0	85.6	91.0	99.1	94.0	87.5	96.8	88.5		91.6

TABLE I: CelebA [2] test set accuracy for each individual attribute and the average accuracy across all attributes. Color indicates rank of method from red (worst) to green (best) – for each attribute individually. Trivial refers to the trivial classifier that always votes for the majority class. The results of the state-of-the-art approaches are obtained from the respected publications. See Sec. III for a detailed discussion of the results.

tion showed that using architectures for resource limited applications can perform almost as good as current top performing architectures. Our methods are fast, accurate, and easy to implement.

We also discussed the contributions of previous works. Especially the results by LNets+ANet [2] show interesting ideas for future research towards a complete end-to-end training e.g. with SSD [26] or Yolo [5]. Also, works by Wang *et al*. [14] and Han *et al*. [1] have shown that multi-task learning can improve face attribute detection performance. They showed that including face recognition or other face related tasks into the same network improves face attribute detection as well.

ACKNOWLEDGEMENT

This work has been funded by the Federal Ministry of Education and Research (BMBF), projects 03ZZ0443G, 03ZZ0459C, and 03ZZ0470. The sole responsibility for the content lies with the authors.

REFERENCES

[1] Hu Han, Anil K. Jain, Fang Wang, Shiguang Shan, and Xilin Chen, "Heterogeneous Face Attribute Estimation: A Deep Multi-Task Learning Approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, jun 2017.

[2] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep Learning Face Attributes in the Wild," in *2015 IEEE International Conference on Computer Vision (ICCV)*. dec 2015, pp. 3730–3738, IEEE.

[3] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation - IEEE Conference Publication," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, Columbus, OH, USA, 2014.

[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1137—-1149, 2016.

[5] Joseph Redmon and Ali Farhadi, "YOLOv3: An Incremental Improvement," *arXiv*, 2018.

[6] Emily M. Hand and Rama Chellappa, "Attributes for Improved Attributes: A Multi-Task Network Utilizing Implicit and Explicit Relationships for Facial Attribute Classification," in *Thirty-First AAAI Conference on Artificial Intelligence*, apr 2017, pp. 4068—-4074.

[7] Davis E. King, "Easily Create High Quality Object Detectors with Deep Learning," 2016.

[8] Davis E. King, "Real-Time Face Pose Estimation," 2014.

[9] Github Repository, "Slim Nasnet," 2018.

[10] Github Repository, "Slim Mobilenet," 2018.

[11] Yang Zhong, Josephine Sullivan, and Haibo Li, "Leveraging mid-level deep representations for predicting face attributes in the wild," in *IEEE International Conference on Image Processing (ICIP)*. sep 2016, pp. 3239–3243, IEEE.

[12] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "FaceNet:

A Unified Embedding for Face Recognition and Clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. jun 2015, pp. 815–823, IEEE.

[13] Manuel Günther, Andras Rozsa, and Terranee E. Boult, "AFFACT: Alignment-free facial attribute classification technique," in *IEEE International Joint Conference on Biometrics (IJCB)*. oct 2017, pp. 90–99, IEEE.

[14] Fang Wang, Hu Han, Shiguang Shan, and Xilin Chen, "Deep Multi-Task Learning for Joint Prediction of Heterogeneous Face Attributes," in *2017 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2017)*. may 2017, pp. 173–179, IEEE.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. 2012, pp. 1097–1105, Curran Associates Inc.

[16] Upal Mahbub, Sayantan Sarkar, and Rama Chellappa, "Segment-based Methods for Facial Attribute Detection from Partial Faces," *arXiv*, 2018.

[17] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit, "Deep Learning is Robust to Massive Label Noise," in *International Conference on Learning Representations (ICLR)*, may 2018.

[18] Barret Zoph, Google Brain, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le, "Learning Transferable Architectures for Scalable Image Recognition," *arXiv*, 2017.

[19] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *arXiv*, jan 2018.

[20] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, dec 2015.

[21] Ilya Loshchilov and Frank Hutter, "SGDR: STOCHASTIC GRADIENT DESCENT WITH WARM RESTARTS," in *5th International Conference on Learning Representations*, Palais des Congrès Neptune, Toulon, France, 2017.

[22] Davis E. King, "Automatic Learning Rate Scheduling That Really Works," 2018.

[23] Andrew G. Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *arXiv*, vol. 1704.04861, apr 2017.

[24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*. 2014, pp. 740–755, Springer, Cham.

[25] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, vol. 111, no. 1, pp. 98–136, jan 2015.

[26] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg, "SSD: Single Shot MultiBox Detector," in *European Conference on Computer Vision (ECCV)*. 2016, pp. 21–37, Springer, Cham.