

ORIGINAL ARTICLE

Computational fairness in adaptive neural networks

Sam Leroux¹  · Ciem Cornelissen¹ · Vishisht Sharma¹ · Pieter Simoens¹

Received: 4 December 2024 / Accepted: 30 May 2025 / Published online: 30 June 2025

© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2025

Abstract

Adaptive neural networks offer a promising approach to reduce the computational cost of neural network inference. These models dynamically allocate computational resources based on each individual input, leading to a significant reduction in average computational cost. In this paper, we show that adaptive computation can have unintended consequences, potentially leading to a significantly worse user experience for certain groups. We train several state-of-the-art adaptive neural networks for the task of age prediction on the FairFace dataset and demonstrate that the average computational cost varies considerably across different user demographics, including age, race and gender. Additionally, we analyze the properties of images that influence the computational cost and show that these factors partially explain the observed demographic differences in performance.

Keywords Bias and fairness · Adaptive neural network · Dynamic computation

1 Introduction

With the rise of artificial intelligence (AI) models deployed in user facing applications, concerns about fair and ethical behavior are growing. These models have been trained on huge datasets that might contain biases and potentially reflect or even amplify societal prejudices. It has been demonstrated multiple times that the accuracy of AI models can differ widely for different demographics in tasks such as gender recognition [1], skin cancer detection [2] and speech recognition [3]. This accuracy discrepancy is largely caused by the use of biased datasets [4] where different groups are underrepresented. Much work has thus been done in an effort to create more balanced datasets such as the FairFace [5], FACET [6] and Balanced Faces in the Wild (BFW) [7] datasets.

In this work, we investigate how susceptible adaptive neural networks are to these biases. Adaptive neural networks can dynamically allocate more or fewer resources to every individual input sample at runtime. They contain some sort of control logic that decides how much compute is needed to return an accurate result for a given input. Averaged out over a large set of inputs, this results in a much lower computational cost. While there has been a lot of work developing adaptive architectures, most of these only report an average classification accuracy on a benchmark dataset such as ImageNet [8]. In this paper, we apply a wide variety of adaptive neural network models to the task of age prediction on the FairFace dataset. We then investigate the accuracy discrepancies between different subgroups of data samples based on age, gender and race. It has been demonstrated multiple times that the performance of machine learning models can differ widely for different demographics [1–3, 9, 10]. However, to the best of our knowledge, no previous work has investigated how these performance discrepancies manifest themselves in adaptive neural networks.

In addition, we introduce the concept of computational fairness. Since adaptive neural networks have different computational costs for different inputs, they might inadvertently allocate more computational resources to certain user groups over others, leading to disparities in energy consumption and processing efficiency. A mobile



application, for example, that executes an adaptive neural network to process user data might consume significantly more energy for one user group compared to others, resulting in a worse user experience for certain demographics. To the best of our knowledge, this work is the first to address this aspect of computational fairness.

The remainder of this paper is structured as follows. In Sect. 2, we provide the required background on adaptive neural network architectures and the evaluation of bias and fairness in machine learning models. We introduce the benchmark dataset and experimental design in Sect. 3. We then apply different adaptive models to the task of age prediction and investigate the biases of this model in terms of accuracy and computational cost in Sect. 4. Finally, Sect. 5 concludes the paper with a summary of our findings and potential directions for future research.

2 Related work

In this section, we provide a brief overview of the main adaptive neural network architectures and the previous work on the evaluation of bias and fairness in deep neural networks.

2.1 Adaptive neural networks

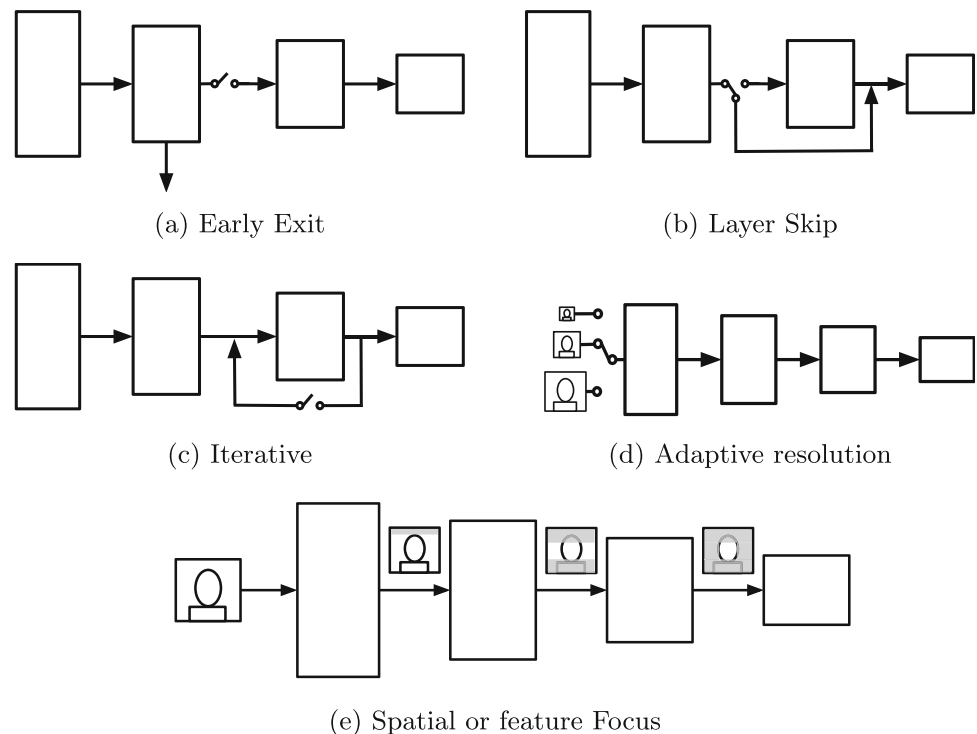
Most commonly used deep neural network architectures have a fixed set of operations that are used to make a prediction based on the input. The computational cost of these models is determined at design time, depending on the number and type of layers in the model. Typically, the architecture is designed to obtain a good overall accuracy on the dataset with an acceptable computational cost and memory footprint. However, a much smaller model would have sufficed for the majority of the input samples. Adaptive neural networks, also sometimes called dynamic neural networks, are able to adjust their computational cost at runtime, depending on the current input. The full capacity of the model remains available for the hardest input samples while easier samples can be automatically processed with a much lower computational cost. Many dynamic architectures have been proposed in the previous years. All of them introduce some sort of decision step into the model architecture that decides at runtime between different computational paths. Figure 1 visually illustrates the most common types of adaptive architectures. We briefly explain these approaches and list the main publications for each technique. We focus on image classification models since these are the most relevant for our work and refer to [11] for a more in-depth overview.

The conceptually easiest to understand family of adaptive neural networks are based on early exits (Fig. 1a). Instead of one output at the end of the network, early exit models have additional exits at intermediate points in the model. While the model is evaluated layer by layer, these exits are able to make intermediate predictions. This idea was first proposed in [12] and was later improved in [13–18]. These methods differ in the model's architecture, the strategy used to determine the exit points and the heuristics that decide to return an early prediction or to continue processing the remaining layers of the model.

A similar idea that also works in the depth dimension of the model is based on skipping layers (Fig. 1b). It is possible to design and train neural networks models in such a way that they are robust to dropping intermediate layers at inference time [19]. This property can be exploited to reduce the computational cost of the model by including a decision mechanism that decides whether a certain layer is needed or not for the current input [20–22]. It is also possible to use a layer multiple times through iterative computation (Fig. 1c). Here, certain parts of the models are repeated, depending on the current input. This was first proposed for recurrent neural networks [23] and later extended to image recognition models [24–26].

Other techniques instead focus on the input data (Fig. 1d). For some inputs, a low resolution version of the input image might already be clear enough to make an accurate prediction. Different approaches have been

Fig. 1 Different strategies to incorporate adaptivity into a deep neural network.



proposed that first process a downscaled version of the input with a very low computational cost and then gradually increase the resolution if the model is not confident enough in its prediction [27–30].

A final, more fine-grained approach learns to focus on specific parts of the input or intermediate feature representations (Fig. 1e). These models learn to ignore the background that is not relevant for the specific task [31, 32] or learn to selectively drop certain feature dimensions that contain less relevant information [33, 34].

These five adaptivity strategies cover the majority of the proposed approaches. Adaptive computation is however an active field of research where new techniques are proposed regularly. Some of these combine two or more of these strategies. RAPNet [35], for example, combines an early exit method with adaptive resolution. Other, more exotic ideas do not necessarily belong to one of these categories. A notable example is dynamic quantization where different quantization levels are used for different inputs [36, 37].

While adaptive neural networks have been shown to reduce the average computational cost, they also have some disadvantages. Not all approaches result in a real-world speed-up. The added control logic can have a severe impact on the throughput on commercially available GPU systems that are optimized for large, dense operations and hence result in a higher inference time. It is therefore important to benchmark the model on the target platform to determine what performance gain can be expected from the adaptive model. A second disadvantage is that adaptive neural networks are more difficult to implement and might not be supported by all frameworks, especially not when targeting edge devices.

Most literature in the field of adaptive neural networks focuses on the trade-off between computational efficiency and model performance. Recently however, a few works have analyzed other properties such as their robustness against adversarial examples [38] or dataset poisoning [39]. Our goal is to extend this line of work with an investigation of bias and fairness in these adaptive architectures.

2.2 Model bias and fairness

Bias and fairness in machine learning (ML) are widely recognized to be critical areas of research due to growing concerns about the societal impact of automated decision-making systems. Bias in ML can arise from various

sources, including biased training data, algorithmic design and the deployment context [40, 41]. These can result in severe accuracy disparities across race and gender in tasks such as facial recognition [9], gender classification [1], emotion recognition [10], skin cancer detection [2] and speech recognition [3].

Several metrics and definitions have been proposed in order to formalize the idea of fairness in ML. In this work, we focus on group fairness which states that different demographic groups should receive similar outcomes [42]. In the context of our work, this would require that different subgroups of people, based on race, gender or age all have a similar model accuracy and computational cost. Other commonly used definitions include individual fairness which ensures that similar individuals are treated similarly [43] or equalized odds, requiring that the probability of a positive outcome should be independent of a protected attribute [44]. Computational fairness can be seen as complementary to existing fairness notions. It extends fairness considerations from “what the model predicts” to “how the model arrives at its predictions” with important implications for user experience, especially in resource-constrained deployment scenarios.

To the best of our knowledge, no other work has explored the presence of bias in adaptive neural networks. Several works however have investigated this for other efficient architectures that use pruning [45] or quantization [46]. While other focus on accuracy discrepancies in these models, we introduce the novel idea of computational fairness where we analyze how the computational cost differs for different user demographics.

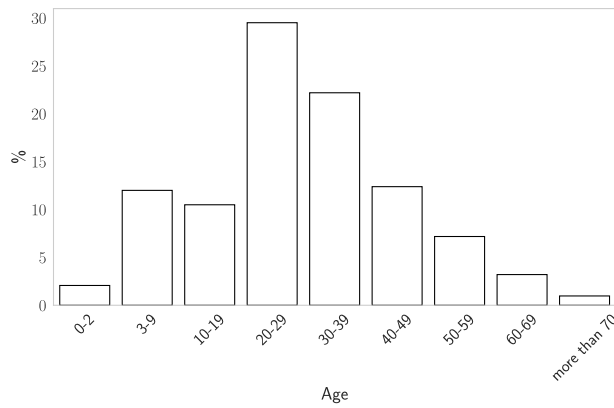
3 Materials and methods

Our goal is to train different adaptive architectures for the age prediction task on the FairFace dataset. We will then quantify the differences in accuracy and computational cost over different demographics. In this section, we provide the details of the training data, model architectures and metrics used to perform this analysis.

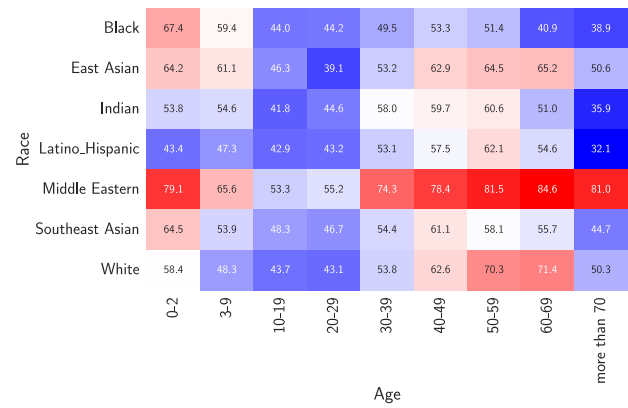
3.1 The FairFace dataset

While there are many face image datasets available, most of them are strongly biased toward Caucasian faces. Models trained on this data will suffer from inconsistent classification accuracy for other race groups, resulting in an unfair system according to the group fairness definition of Sect. 2.2. The FairFace dataset [5] attempts to address this issue. It contains 108,501 images, balanced over 7 race groups: White, Black, Indian, East-Asian, Southeast Asian, Middle Eastern and Latino. In addition, each image is labeled with a gender (male/female) and age group (0–2, 3–9, 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70+). It is important to mention that while the dataset is more-or-less balanced over race and gender, it is heavily skewed along the age dimension with almost a third of the data belonging to the 20–29 group. Figure 2a shows this distribution of samples over different age groups. The two smallest groups 0–2 and 70+ only contain two and one percent of the data, respectively. There are also more subtle imbalances in this dataset. Figure 2b shows for each age–race combination the fraction of males in that subgroup. We can see large discrepancies over different subgroups. For the Middle Eastern group, for example, 85% of the data samples in the 60–69 group are male while this is only the case for 32% of the samples in the Latino 70+ group. This example illustrates that even datasets that were specifically collected to avoid imbalances might still contain subtle biases that can influence models trained on them.

In all our experiments, we followed the original train–test split. We used the version of the dataset with padding = 1.25 which contains more background information. This makes the task slightly more challenging but also offers opportunities for adaptive neural networks to optimize their computational cost. The original age labels were grouped in ten-year increments, which we found to be somewhat arbitrary. Therefore, we adopted a more meaningful categorization, similar to the approach in [47], and reorganized the data into five semantically relevant groups: 0–9 (children), 10–19 (teenagers), 20–39 (adults), 40–59 (middle-aged) and 60+ (seniors). We did not balance the dataset in any other way since the main goal of this paper is to investigate how biases in the dataset influence the behavior of adaptive neural networks.



(a)



(b)

Fig. 2 Despite the FairFace dataset being more-or-less balanced over different races and genders, there is still a large imbalance over different ages (a). We also found significant differences for age–race subgroups (b). In this figure, white entries are more-or-less balanced over gender. Blue entries are predominantly female while red entries contain mostly data from males.

3.2 Models

Table 1 lists the models used in our experiments, including nine adaptive architectures selected to represent the full spectrum of adaptivity strategies discussed in Sect. 2. We evaluate both convolutional architectures and vision transformer-based models. Additionally, two non-adaptive architectures are included as baselines: a ResNet50 model [48] and a vision transformer (ViT-base-patch16-224) [49]. For reference, we provide each model’s top-1 accuracy on the ImageNet dataset, along with the total number of parameters.

All models were implemented in PyTorch, using either publicly available source code or code provided by the original authors. Each model was trained from scratch, following the original training procedures and hyperparameters used for ImageNet, which we found to generalize well to the age prediction task. However, it is worth noting that more fine-tuned hyperparameter adjustments could likely yield slightly higher accuracy for some models. We resized input images to 224×224 , consistent with ImageNet standards, and applied random horizontal flips during training as data augmentation.

Table 1 Model details

Model	Adaptivity type	ImageNet top-1 (%)	Parameters
ResNet50 [48]	n/a	79.3	23.5 M
Vit [49]	n/a	81.1	85.8 M
IamNN [25]	Iterative	69.5	5.4 M
Dynamic ViT [50]	Adaptive resolution	81.5	21.7 M
Cascade (ResNet18 base) [12]	Early exit	68.7	11.1 M
ConvNet-AIG [20]	Layer skip	76.2	23.7 M
IMTA [17]	Early exit	72.3	22.0 M
CF-ViT [51]	Adaptive resolution	80.7	22.8 M
RANet [27]	Adaptive resolution	74.3	116.2 M
SkipNet [21]	Layer skip	77.5	22.0 M
DynConv [52]	Spatial	81.2	23.7 M

3.3 Metrics

We analyze two aspects of the models: their accuracy and their computational cost. While most works focus only on the total accuracy and the average computational cost over the entire dataset, we are less interested in these exact values and instead will focus on their differences for different demographics.

3.3.1 Accuracy

To test whether the accuracy obtained by the model is significantly different across various demographics based on age, gender or race, we report the Chi-square test of independence. This test evaluates whether the distribution of correct and incorrect predictions (i.e., the model's accuracy) varies significantly across different demographic groups. For this test, our null hypothesis is that there is no association between the sensitive attribute and the model correctness. The Chi-square test results in a χ^2 statistic measuring the strength of the association and a p-value that represents the probability of observing the data assuming the null hypothesis is true. (There is no association between this attribute and the model accuracy.) We use a threshold of $p < 0.05$ to reject the null hypothesis (i.e., conclude that there is a significant association between sensitive attribute and model correctness). For all experiments, we report the χ^2 statistic to quantify the strength of the association.

3.3.2 Computational cost

To quantify the computational cost of the model, we measure the number of floating-point operations (FLOPs) for a single forward pass through the model. It should be noted that this can be misleading in the sense that a reduction in the number of FLOPs does not always result in a real-world improved run time of the model. This discrepancy arises because actual runtime performance is influenced by various factors beyond FLOPs, such as memory access patterns, hardware-specific optimizations, parallelism and data movement overhead. Consequently, even if a model has fewer FLOPs, it may not necessarily run faster in practice, especially on different architectures or under varying system loads. We still opted to use FLOPs to quantify the model's computational cost since it is an objective metric, independent of implementation and system architecture. In addition, it is used in almost all related works.

To measure how different the computational cost of the models is over different groups, we perform an analysis of variance test (ANOVA). This test results in an F-statistic where a higher value indicates more variance in computational cost between groups relative to the variance within the groups. Similar to the Chi-square test from previous section, we define the null hypothesis as follows: The mean computational cost is the same across all groups (i.e., no significant difference in computational cost based on demographic categories such as age, gender or race). If the ANOVA yields a high F-statistic and a low p-value ($p < 0.05$), we reject the null hypothesis, meaning that the computational cost varies significantly between the groups. For all experiments, we report the F-statistic to quantify the strength of the association.

4 Results

4.1 Model performance

We first evaluate the accuracy of the different models on the age prediction task. Table 2 summarizes for each model the total overall accuracy on this task, together with the χ^2 statistic. All models obtain a similar overall accuracy between 75% and 79%. For the race attribute, we find a significant difference in accuracy for all models, including the two baseline models. The accuracy for the Black group was consistently lowest for all models, while that of the East-Asian group was always the highest. The gap between the two was the largest for the ViT

Table 2 Overall age prediction accuracy on the FairFace dataset for the different models, together with the coefficient of variation in accuracy over different race, gender and age groups

Model	Accuracy	χ^2_{race}	χ^2_{gender}	χ^2_{age}	Balanced accuracy
IamNN	75.91 %	19.67		2413.51	63.08 %
DynamicViT	77.34 %	34.64		2114.24	67.41 %
Cascade	76.65 %	19.69		2763.38	63.27 %
ConvNet-AIG	75.71 %	16.37		1880.91	64.88 %
IMTA	74.51 %	22.37		1853.48	63.78 %
CF-ViT	77.36 %	21.16		2160.03	66.19 %
RANet	75.33 %	25.19		2720.45	62.57 %
SkipNet	76.61 %	28.06		1405.87	68.81 %
DynConv	76.81 %	19.91	4.85	1712.59	67.12 %
Vit	77.44 %	39.02		2140.52	65.03 %
ResNet50	79.04 %	20.28		1503.37	70.74 %

baseline model (82.9% vs. 74.1%). The reason for this accuracy discrepancy is unclear. Both groups have a similar number of data points and a similar distribution over age groups suggesting that the data quality or image properties (low brightness, blurriness) might be a problem. For the gender attribute, only the DynConv model obtained a significant difference in model accuracy for male and female persons (77.9% vs. 75.8%). Finally, for the age attribute, we observe very large accuracy differences for different groups where group 0–9 and 20–39 obtained the highest accuracies and group 10–19 consistently the lowest. Part of this can be explained by the large amount of training data for the 20–39 group. For all sensitive attributes, the χ^2 statistic for the adaptive models is similar to that of the baseline models suggesting that the adaptive nature of these models does not result in significantly higher accuracy discrepancies between groups. For completeness, we also report the balanced accuracy (the average recall over each class) in the last column.

4.2 Computational cost

Table 3 presents the average computational cost for each model across all test samples, along with the corresponding standard deviation. A larger standard deviation suggests that the model is more adaptive, meaning the computational cost varies more between samples. Since the different architectures have vastly different

Table 3 Average computational cost (FLOPs) for each model, together with its standard deviation and the coefficient of variation. For each model, we show the ANOVA F-metric as a measure of how the mean computational cost differs between groups formed based on race, gender and age. A higher F-metric indicates larger differences in mean computational cost. Results that were not statistically significant ($p > 0.05$) have been omitted from this table

Model	AVG	STD	CV	F_{race}	F_{gender}	F_{age}
IamNN	2.49	0.14	0.06	13.98	20.35	50.41
DynamicViT	1.08	1.05	0.97	12.14		103.01
Cascade	1.59	0.37	0.23	21.57		409.56
ConvNet-AIG	3.26	0.30	0.09	5.87		134.37
IMTA	0.54	0.33	0.63	10.03		72.65
CF-ViT	8.54	1.21	0.14	10.80	32.58	321.28
RANet	0.62	0.43	0.69	38.22	21.65	385.34
SkipNet	3.44	0.19	0.06	33.47		160.76
DynConv	3.18	0.13	0.04	15.80	26.17	41.45
Vit	16.86	0.00	0.00	n/a	n/a	n/a
ResNet	4.13	0.00	0.00	n/a	n/a	n/a

computational costs, we cannot directly compare the standard deviations between models. Therefore, the coefficient of variation ($CV = \text{std}/\text{avg}$) is provided in the third column which allows for more meaningful comparisons. The computational cost of the baseline methods (ViT/ResNet) is constant, hence a standard deviation of zero.

We then analyze the differences in average computational cost for groups based on race, gender and age. We report the ANOVA F-statistic as described in Sect. 3.3. A higher value indicates a larger discrepancy in computational cost for different groups. For clarity, we omitted results that were not statistically significant ($p > 0.05$). Similar to the results of Sect. 4.1, we notice the largest differences for the age attribute. This can probably be attributed to the large imbalance in training data for different age groups. For the remainder of the experiments, we therefore focus on the race and gender attributes.

For the race attribute, all models have significantly different computational costs for different groups. We find that images from the Black group consistently result in the highest computational cost. This is not surprising as this group also had the lowest accuracy which causes the models to learn that more resources are needed to obtain an accurate prediction. The group with the lowest cost differs between models but is typically either “East-Asian,” “Middle Eastern” or “White.” We find that the RANet model results in the largest F-statistic for the race attribute. For this model, the average computational cost for the Black group is 40% higher than that of the East-Asian group. It is hard to pinpoint the underlying reason for this difference. Both groups have the same amount of training samples and a similar distribution over age and gender categories which suggests that this difference cannot simply be attributed to quantitative differences in the dataset. It does however highlight the importance of this analysis. This difference in computational cost could have real-world implications on the perceived user experience for different group. A mobile application, for example, that evaluates this model on user data would drain the battery of the device much faster for users with certain racial backgrounds compared to others.

To gain some insight into the behavior of the models, we show a few randomly selected test images with a low and a high computational cost for each model in Fig. 4. We have to be careful not to draw strong conclusions from these but a few patterns do stand out. For example, for the IamNN model, we notice that the images with a low computational cost all show centered faces that look straight ahead into the camera (annotated with a green border). Images with the highest computational cost on the other hand typically depict people looking to the side (annotated with a red border). In an attempt to find these underlying properties of the images that influence the computational cost, we extracted ten features that we suspect might have an influence:

- **Brightness:** A weighted sum of the RGB values, based on how the human eye perceives different colors: $\text{Perceived Brightness} = 0.299 \times R + 0.587 \times G + 0.114 \times B$. This formula is based on ITU-R BT.601 recommendation [53]. A higher value indicates a higher brightness (i.e., more white). We hypothesize that images with a lower brightness are more difficult to process and hence will result in a larger computational cost.
- **Sharpness:** We calculate the variance of the Laplacian to capture edges and high-frequency components. A higher value indicates a sharper image. We hypothesize that a blurry image will result in a larger computational cost.
- **Profileness:** This metric has a high value when the person is looking to the side (profile) and a low value when he is looking straight ahead into the camera. We used the RetinaFace [54] model to extract the eye and nose positions from each image and used this to determine the side-to-side position of the face. Similar to the example given previously, we hypothesize that images where the person is looking straight into the camera are easier to predict.
- **Area:** The relative portion of the image occupied by the main face. We used the RetinaFace model to predict a bounding box around the face. We hypothesize that a larger area provides more details that are useful for age prediction.

- **Number of faces:** The total number of faces detected in the image by the RetinaFace model. While each image is labeled with a single age, some images also contain faces in the background. We hypothesize that this might confuse the model, resulting in a higher computational cost.
- **Contrast:** RMS contrast defined as the standard deviation of the grayscale pixel intensities. Similar to the brightness, we hypothesize that a lower contrast has a negative impact on model performance and hence will result in a higher computational cost.
- **Color:** The Hasler–Susstrunk colorfulness metric [55]. A high value indicates a more vibrant colorful image.
- **Edge density:** We run the Canny edge detection algorithm on the image and calculate the ratio of the number of edge pixels to the total number of pixels. This metric can be used as a proxy for “busyness” of the image.
- **Entropy:** The Shannon entropy of the pixel intensities. An image with a lot of different textures and objects will have higher entropy.
- **Symmetry:** as measured by the Structural Similarity Index Measure (SSIM) [56] between the left half of the image and the right half.

We extracted these metrics for every test sample and calculated the Pearson correlation coefficient between these metrics and the computational cost of each model. Table 4 shows these results. A positive value indicates that the computational cost of this model tends to increase as the feature increases. A negative value indicates the opposite. For clarity, we only show the results that were statistically significant ($p < 0.05$). We find that most models are sensitive to the “profileness” of the picture, i.e., images where the person is looking to the side have a higher computational cost. This is the most obvious for the IamNN model, which confirms our previous observation. The only model that is not influenced by this is the DynConv architecture. The performance of this model however has a large negative correlation with the colorfulness of the image. This can also be noted in the examples of Fig. 4 where all the images with the lowest computational cost seem to be dominated by a blue or green hue (annotated with an orange border). The area occupied by the face is an interesting property. For some models (IamNN, DynConv), the computational cost increases as a larger part of the image is occupied by the face while for other models (SkipNet, ConvNet-AIG), the opposite can be observed. This behavior can probably attributed to the different adaptive strategies used by these models. The DynConv model uses spatial adaptivity where different regions of the image are processed with different amounts of resources. As the face occupies a larger part of the image, a larger part will need to be processed and hence, the computational cost will increase. The SkipNet and ConvNet-AIG models on the other hand implement a layer skip strategy. Images where a single, large face is clearly visible in the center of the frame might require fewer layers to extract the necessary features to make an accurate prediction.

Table 4 Pearson correlation coefficient between image feature values and the computational cost of the adaptive architectures. A positive value indicates that the computational cost of this model tends to increase as the feature increases. A negative value indicates the opposite. Empty entries were not found to be statistically significant (i.e., $p > 0.05$)

Model	Brightness	Sharpness	Profile	Area	Num	Contrast	Color	Edges	Entropy	Symmetry
IamNN		−0.06	0.37	0.25	−0.06	−0.04	−0.09	−0.12	−0.03	−0.05
DynamicViT			0.05		0.02			0.01		
Cascade		−0.03	0.15			−0.03	0.04		−0.03	−0.01
ConvNet-AIG			0.07	−0.06						
IMTA			0.06		0.03			−0.02		
CF-ViT		−0.03	0.12	0.08	0.03	−0.02	0.02	−0.04	−0.02	−0.02
RANet	0.03	−0.02	0.15	0.08	0.03			−0.02		
SkipNet	−0.10	−0.05	0.16	−0.21	−0.11	−0.18	0.04	−0.09	−0.07	0.09
DynConv	0.09	−0.09		0.12	0.10	−0.11	−0.28		−0.11	−0.03

We now use these insights in an attempt to explain some of the observed biases. We analyze how different the features are on average for images belonging to different groups based on race, gender and age. Table 5 shows the ANOVA F-metric for each attribute/feature combination. A higher value indicates a larger difference in feature value on average for different groups based on the sensitive attribute. As in the previous results, we only show entries that are statistically significant ($p < 0.05$). By combining these results with those of Table 4, we can attempt to explain some of the differences in computational cost of Table 3. For example, the strong link between the “colorfulness” feature and gender combined with the sensitivity of the DynConv model to this feature might explain the differences in computational cost of this model for male and female images.

While we believe that this approach can help to gain some insight into the behavior of the adaptive models and hence might partially explain the observed biases. We should be careful not to draw strong conclusions as the true underlying causes are most likely much more complex and based on interactions between different properties of the test image and training data. It is easy to find examples where our approach is unable to explain the observed behavior. For example, we find that the Cascade model exhibits the most significant computational bias for the age attribute (Table 3). Using the results of Table 5, we know that age is most strongly correlated with image brightness and face area; however, neither of these seem to have a significant correlation with the Cascade model’s computational cost (Table 4) and hence cannot explain the computational (un)fairness.

It is important to note that in all our results, we analyzed one sensitive attribute at a time. This means that we calculate aggregated statistics over the other attributes. When analyzing the “race” attribute for example, we do not take the gender and age into account. This means that we will ignore intra-group differences. To analyze this, we show the average computational cost for every race–age subgroup in Figs. 3 and 4 clearly see large differences within the same age or race group. For example, for the “Indian” group, we find that samples from the “10–19” age group result in a significantly higher computational cost compared to other age groups for the IMTA model while the exact opposite is true for the Cascade model. A similar observation can be made for the “White,” “0–9” group in the IamNN and SkipNet model. This suggests that a more in-depth analysis would be needed to fully uncover the causes of the observed biases. As we divide the dataset into more detailed subgroups however, we quickly run into the limitations of the dataset. In the FairFace validation set for example, there are only eight images of Middle Eastern 70+ women and only six images of Black children between the ages of 0 and 2. A much larger (balanced) dataset would be needed to draw more detailed conclusions.

Table 5 ANOVA F-metric for feature values over different groups based on race, gender or age. Results that were not statistically significant ($p > 0.05$) have been omitted from this table

Feature	F_{race}	F_{gender}	F_{age}
Brightness	5.65	23.57	25.47
Sharpness	4.32	25.08	6.88
Profile	5.88	33.60	14.38
Area	12.84	8.86	22.81
Number of faces	11.57		12.26
Contrast	9.86		3.05
Color	2.71	78.02	3.42
Edges	3.75	21.89	16.20
Entropy	5.90	4.70	
Symmetry	6.03	63.63	5.56

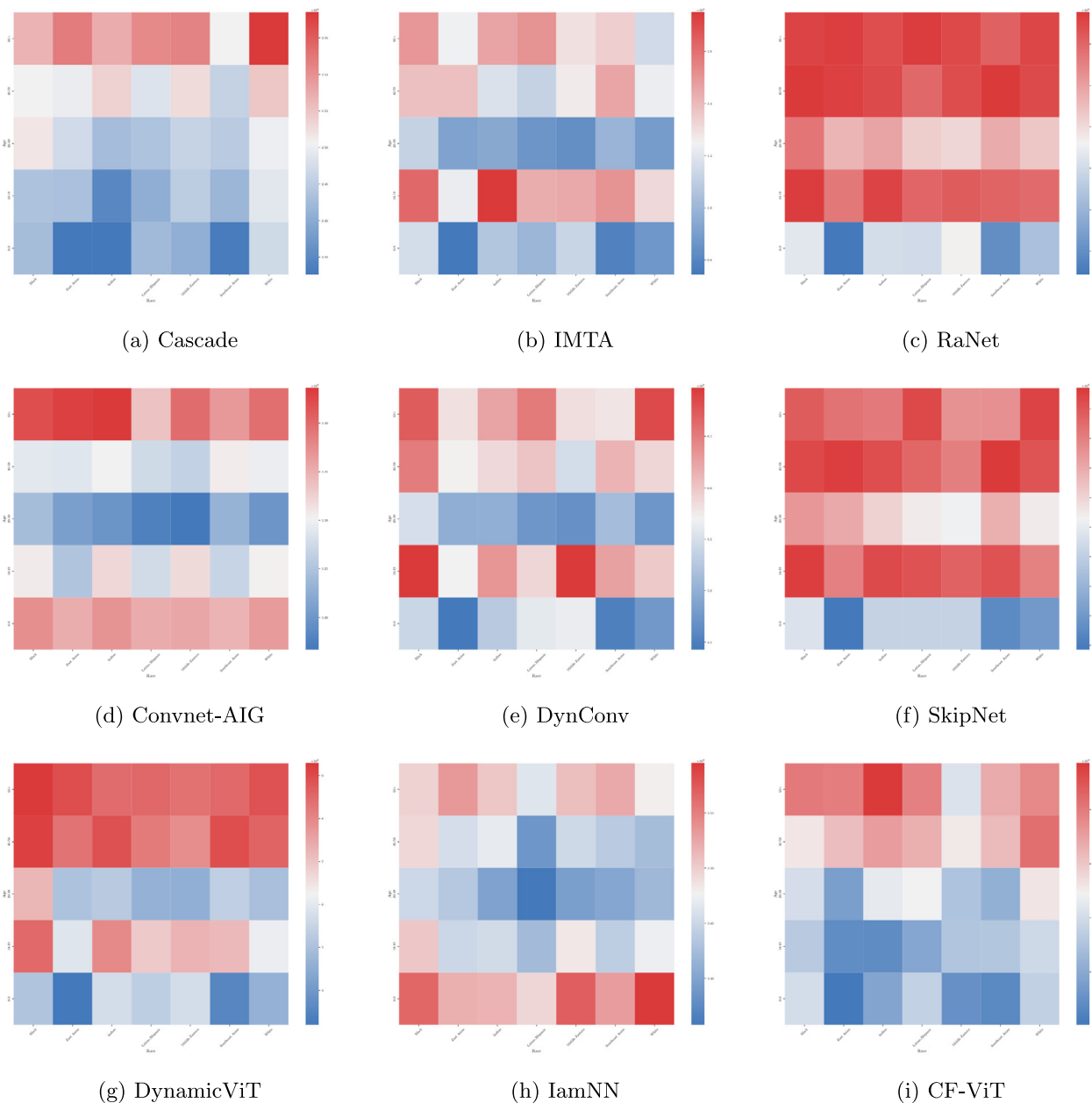


Fig. 3 Average computational cost of each model for different age–race subgroups. A blue value indicates a lower cost, while a red value indicates a higher computational cost. These colors are normalized for each model independently and cannot be used directly to compare computational costs between models.

5 Conclusion and future work

Adaptive architectures are a promising approach to reduce the inference cost of deep neural networks. When these models are deployed in user facing applications however, we need to go beyond calculating an average accuracy and computational cost as these can hide significant differences for different user demographics. In this paper, we found that all tested adaptive architectures result in statistically significantly different computational costs for images sourced from different age, gender and race groups. These differences could result in a lower user

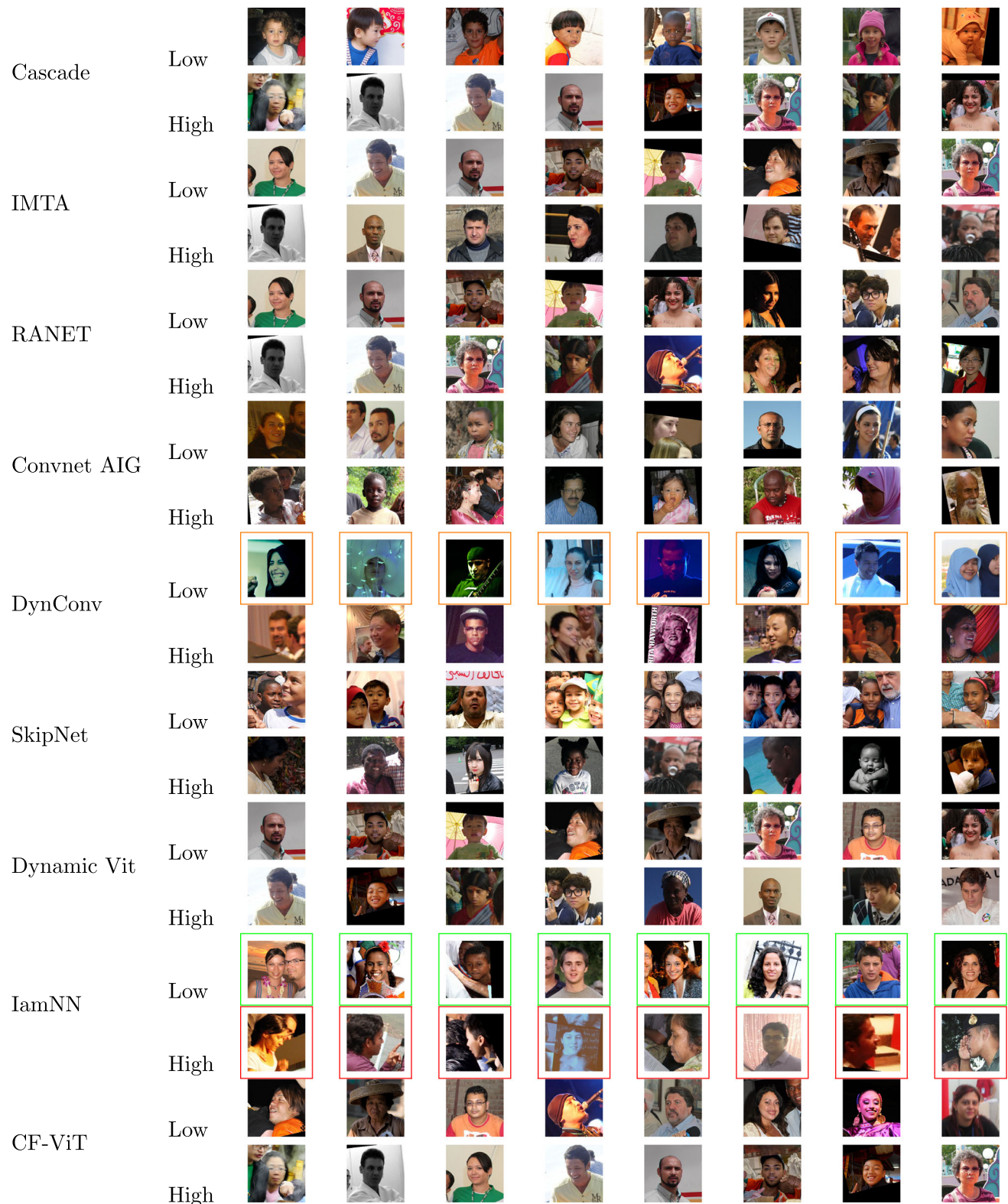


Fig. 4 Examples of test images with a low and high computational cost for the different adaptive neural networks.

experience or increased cost for certain subgroups. We analyzed the sensitivity of the different models to image features such as brightness or head pose and were able to explain some of these differences.

More work is needed to understand the significance and underlying causes of these observations. There are several limitations to the results presented in this work. First of all, the dataset used is relatively small, especially when conclusions are being made for combinations of different attributes. Secondly, while the FairFace dataset was carefully collected to avoid stereotypical images and similar biases, it is still a dataset sourced from publicly available images on the internet, annotated by humans. It is more than likely that there are some subtle biases introduced this way. Finally, as adaptive neural network architectures are black box methods, it is unclear what aspects influence the computational cost. The low level image attributes used in this work are able to partially explain some of the differences but the underlying causes will be much more complex. We hope that this work encourages developers of adaptive architectures to take the aspect of computational fairness into account. Some interesting areas for follow-up work would be to

- Use interpretable AI algorithms to uncover more underlying causes of computational bias.
- Design metrics to measure and compare the computational bias of different models.
- Design cost functions and regularization approaches to avoid computational bias.
- Collect larger, balanced datasets annotated with multiple demographic attributes.

Acknowledgements This research was supported by funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program.

Author contributions Sam Leroux contributed to conceptualization, software and writing.
Ciem Cornelissen and Vishist Sharma contributed to conceptualization, review and editing.
Pieter Simoens helped in project administration.

Funding This research was supported by funding from the Flemish Government under the “Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen” program.

Data availability The FairFace dataset used in this work is publicly available at <https://github.com/joojs/fairface> Source code can be made available upon request.

Declarations

Conflict of interest The author(s) declare(s) that they have no conflict of interest.

References

1. Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. PMLRpp, pp 77–91
2. Guo LN, Lee MS, Kassamali B, Mita C, Nambudiri VE (2022) Bias in, bias out: underreporting and underrepresentation of diverse skin types in machine learning research for skin cancer detection-a scoping review. *J Am Acad Dermatol* 87(1):157–159
3. Martin JL, Wright KE (2023) Bias in automatic speech recognition: the case of African American language. *Appl Linguis* 44(4):613–630
4. Fabbri S, Papadopoulos S, Ntoutsis E, Kompatsiaris I (2022) A survey on bias in visual datasets. *Comput Vis Image Underst* 223:103552
5. Karkkainen K, Joo J (2021) FairFace: face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 1548–1558
6. Gustafson L, Rolland C, Ravi N, Duval Q, Adcock A, Fu C-Y, Hall M, Ross C (2023) Facet: fairness in computer vision evaluation benchmark. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 20370–20382
7. Robinson JP, Qin C, Henon Y, Timoner S, Fu Y (2023) Balancing biases and preserving privacy on balanced faces in the wild. *IEEE Trans Image Process* 32:4365–4377
8. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. *Int J Comput Vis* 115(3):211–252. <https://doi.org/10.1007/s11263-015-0816-y>

9. Nagpal S, Singh M, Singh R, Vatsa M (2019) Deep learning for face recognition: pride or prejudiced? arXiv preprint [arXiv:1904.01219](https://arxiv.org/abs/1904.01219)
10. Howard A, Zhang C, Horvitz E (2017) Addressing bias in machine learning algorithms: a pilot study on emotion recognition for intelligent systems. In: 2017 IEEE workshop on advanced robotics and its social impacts (ARSO). IEEE, pp 1–7
11. Han Y, Huang G, Song S, Yang L, Wang H, Wang Y (2021) Dynamic neural networks: a survey. *IEEE Trans Pattern Anal Mach Intell* 44(11):7436–7456
12. Leroux S, Bohez S, Verbelen T, Vankeirsbilck B, Simoens P, Dhoedt B (2015) Resource-constrained classification using a cascade of neural network layers. In: 2015 international joint conference on neural networks (IJCNN). IEEE, pp 1–7
13. Leroux S, Bohez S, De Coninck E, Verbelen T, Vankeirsbilck B, Simoens P, Dhoedt B (2017) The cascading neural network: building the internet of smart things. *Knowl Inf Syst* 52:791–814
14. Teerapittayanon S, McDanel B, Kung H-T (2016) BranchyNet: fast inference via early exiting from deep neural networks. In: 2016 23rd international conference on pattern recognition (ICPR). IEEE, pp 2464–2469
15. Bolukbasi T, Wang J, Dekel O, Saligrama V (2017) Adaptive neural networks for efficient inference. In: International conference on machine learning. PMLR, pp 527–536
16. Wang M, Mo J, Lin J, Wang Z, Du L (2019) DynExit: a dynamic early-exit strategy for deep residual networks. In: 2019 IEEE international workshop on signal processing systems (SiPS). IEEE, pp 178–183
17. Li H, Zhang H, Qi X, Yang R, Huang G (2019) Improved techniques for training adaptive deep networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1891–1900
18. Passalis N, Raitoharju J, Tefas A, Gabbouj M (2020) Efficient adaptive inference for deep convolutional neural networks using hierarchical early exits. *Pattern Recogn* 105:107346
19. Greff K, Srivastava RK, Schmidhuber J (2016) Highway and residual networks learn unrolled iterative estimation. arXiv preprint [arXiv:1612.07771](https://arxiv.org/abs/1612.07771)
20. Veit A, Belongie S (2018) Convolutional networks with adaptive inference graphs. In: Proceedings of the European conference on computer vision (ECCV), pp 3–18
21. Wang X, Yu F, Dou Z-Y, Darrell T, Gonzalez JE (2018) SkipNet: learning dynamic routing in convolutional networks. In: Proceedings of the European conference on computer vision (ECCV), pp 409–424
22. Wu Z, Nagarajan T, Kumar A, Rennie S, Davis LS, Grauman K, Feris R (2018) BlockDrop: dynamic inference paths in residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8817–8826
23. Graves A (2016) Adaptive computation time for recurrent neural networks. arXiv preprint [arXiv:1603.08983](https://arxiv.org/abs/1603.08983)
24. Guo Q, Yu Z, Wu Y, Liang D, Qin H, Yan J (2019) Dynamic recursive neural network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 5147–5156
25. Leroux S, Molchanov P, Simoens P, Dhoedt B, Breuel T, Kautz J (2018) IamNN: iterative and adaptive mobile neural network for efficient image classification. arXiv preprint [arXiv:1804.10123](https://arxiv.org/abs/1804.10123)
26. Figurnov M, Collins MD, Zhu Y, Zhang L, Huang J, Vetrov D, Salakhutdinov R (2017) Spatially adaptive computation time for residual networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1039–1048
27. Yang L, Han Y, Chen X, Song S, Dai J, Huang G (2020) Resolution adaptive networks for efficient inference. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2369–2378
28. Wang Y, Sun F, Li D, Yao A (2020) Resolution switchable networks for runtime efficient image recognition. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XV 16. Springer, pp 533–549
29. Wieme J, Mortier R, Leroux S, Cool S, Pieters J, Maes W. Efficient detection of colorado potato beetles in ultra-high-resolution UAV-imaging of potato fields using convolutional neural networks
30. Zhu M, Han K, Wu E, Zhang Q, Nie Y, Lan Z, Wang Y (2021) Dynamic resolution network. *Adv Neural Inf Process Syst* 34:27319–27330
31. Figurnov M, Ibraimova A, Vetrov DP, Kohli P (2016) PerforatedCNNs: acceleration through elimination of redundant convolutions. *Advances in neural information processing systems* 29
32. Xie Z, Zhang Z, Zhu X, Huang G, Lin S (2020) Spatially adaptive inference with stochastic feature sampling and interpolation. In: Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, Part I 16. Springer, pp 531–548
33. Hua W, Zhou Y, De Sa C, Zhang Z, Suh GE (2019) Boosting the performance of CNN accelerators with dynamic fine-grained channel gating. In: Proceedings of the 52nd annual IEEE/ACM international symposium on microarchitecture, pp 139–150
34. Leroux S, Bohez S, De Boom C, De Coninck E, Verbelen T, Vankeirsbilck B, Simoens P, Dhoedt B (2016) Lazy evaluation of convolutional filters. arXiv preprint [arXiv:1605.08543](https://arxiv.org/abs/1605.08543)
35. Hu Y, Cheng Y, Zhou Z, Cao Z, Lu A, Liu J, Zhang M, Li Z (2024) RapNet: resolution-adaptive and predictive early exit network for efficient image recognition. *IEEE Internet of Things J* 11:20

36. Liu Z, Wang Y, Han K, Ma S, Gao W (2022) Instance-aware dynamic neural network quantization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12434–12443
37. Allwin PS, Gomony MD, Geilen M (2024) Run-time non-uniform quantization for dynamic neural networks in wireless communication. In: 2024 29th Asia and South Pacific design automation conference (ASP-DAC). IEEE, pp 915–920
38. Haque M, Yang W (2023) Dynamic neural network is all you need: Understanding the robustness of dynamic mechanisms in neural networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 1497–1506
39. Chen S, Chen H, Haque M, Liu C, Yang W. SlothBomb: efficiency poisoning attack against dynamic neural networks
40. Barocas S, Selbst AD (2016) Big data's disparate impact. *Calif. L. Rev.* 104:671
41. Du M, Yang F, Zou N, Hu X (2020) Fairness in deep learning: a computational perspective. *IEEE Intell Syst* 36(4):25–34
42. Binns R (2020) On the apparent conflict between individual and group fairness. In: Proceedings of the 2020 conference on fairness, accountability, and transparency, pp 514–524
43. Fleisher W (2021) What's fair about individual fairness? In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society, pp 480–490
44. Hardt M, Price E, Srebro N (2016) Equality of opportunity in supervised learning. *Advances in neural information processing systems* 29
45. Iofinova E, Peste A, Alistarh D (2023) Bias in pruned vision models: In-depth analysis and countermeasures. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 24364–24373
46. Slyman E, Kanneganti A, Hong S, Lee S (2024) You never know: quantization induces inconsistent biases in vision-language foundation models. *arXiv preprint [arXiv:2410.20265](https://arxiv.org/abs/2410.20265)*
47. Wu H, Bowyer KW (2023) What should be balanced in a “balanced” face recognition dataset? *arXiv preprint [arXiv:2304.09818](https://arxiv.org/abs/2304.09818)*
48. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
49. Dosovitskiy A (2020) An image is worth 16x16 words: transformers for image recognition at scale. *arXiv preprint [arXiv:2010.11929](https://arxiv.org/abs/2010.11929)*
50. Wang Y, Huang R, Song S, Huang Z, Huang G (2021) Not all images are worth 16x16 words: dynamic transformers for efficient image recognition. *Adv Neural Inf Process Syst* 34:11960–11973
51. Chen M, Lin M, Li K, Shen Y, Wu Y, Chao F, Ji R (2023) Cf-vit: a general coarse-to-fine method for vision transformer. In: Proceedings of the AAAI conference on artificial intelligence, vol 37, pp 7042–7052
52. Verelst T, Tuytelaars T (2020) Dynamic convolutions: exploiting spatial sparsity for faster inference. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 2320–2329
53. BT R, et al (2011) Studio encoding parameters of digital television for standard 4: 3 and wide-screen 16: 9 aspect ratios. In: International radio consultative committee international telecommunication union, Switzerland, CCIR Rep
54. Deng J, Guo J, Zhou Y, Yu J, Kotsia I, Zafeiriou S (2019) Retinaface: Single-stage dense face localisation in the wild. *arXiv preprint [arXiv:1905.00641](https://arxiv.org/abs/1905.00641)*
55. Hasler D, Suesstrunk SE (2003) Measuring colorfulness in natural images. In: *Human Vision and Electronic Imaging VIII*, vol 5007, pp 87–95. SPIE
56. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process* 13(4):600–612

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Authors and Affiliations

Sam Leroux¹  · Ciem Cornelissen¹ · Vishisht Sharma¹ · Pieter Simoens¹

✉ Sam Leroux
sam.leroux@ugent.be

Ciem Cornelissen
ciem.cornelissen@UGent.be

Vishisht Sharma
vishisht.sharma@ugent.be

Pieter Simoens
pieter.simoens@ugent.be

- ¹ IDLab, Department of Information Technology, Ghent University - imec, Technologiepark-Zwijnaarde 126, 9052 Gent, Belgium