Improving Fairness in Visual Recognition through Feature Distillation

CSC 591/791: Deep Learning Beyond Accuracy — Term Paper

Aravinda Jatavallabha (arjatava), Aadithya Naresh (anaresh), and Mihir Arora (marora6)
North Carolina State University, USA

Abstract

Ensuring fairness is very important in computer vision systems, which are used in areas that directly affect people's lives. For example, facial recognition, employment tools, and surveillance often show unfairness against certain demographic groups. This can worsen existing social inequalities. Although researchers are focusing on fairness in machine learning, the solutions often require starting from scratch and retraining models with complex rules to achieve fairness. This process demands a lot of computer resources, making it difficult and impractical, especially for larger systems or those that are already functioning.

In this study, we implement a method called Maximum Mean Discrepancy-based Fair Distillation (MFD) inspired by Jung, S., et al. [1]. This approach helps transfer fair and group-neutral information from a biased "teacher" model to a "student" model. By using MMD, we can make sure that the student's features for each group match the teacher's average features for that group. This process promotes fairness and keeps prediction performance strong. This method is carefully designed to ensure consistent results within each group and good alignment between different groups. We also support our approach with solid theoretical explanations and test it on both made-up and real data. Our findings show that MFD effectively reduces demographic bias without reducing model accuracy. This makes it a practical and reliable solution for using AI fairly.

I. Introduction

Deep neural networks have noticeably improved computer vision technology, leading to big changes in several fields. These fields include facial recognition, automatic surveillance, AI-supported hiring, and medical diagnostics. While these systems offer many benefits, they also pose serious risks. A big concern is bias. If the models are trained with unbalanced data, they might wrongly categorize individuals based on sensitive characteristics like race or gender. This can result in unfair treatment in real-life scenarios. Due to this issue, there is now more focus on making AI systems fair, particularly those involved in important or sensitive decision-making processes.

In machine learning, fairness means ensuring that a model's predictions are not unfairly influenced by factors like race, gender, or age. Achieving this fairness can be challenging, particularly when trying to maintain prediction accuracy. There are three primary strategies to enhance fairness in models:

1) Pre-processing: This involves altering the data before training to eliminate bias.[2][3]
2) In-processing: This technique applies fairness rules during the model's training process. [4][5][6]
3) Post-processing: This approach adjusts the model's predictions after training is complete.

In-processing methods are especially popular because they directly apply fairness rules throughout the model's learning process. However, they often require full access to the training data and models and need significant computational resources to restart the training from scratch.

We implemented a method called Maximum Mean Discrepancy-based Fair Distillation (MFD) to help create fair models more efficiently. Unlike starting from scratch, MFD uses knowledge distillation. This technique typically makes models smaller, but here it helps us choose only the unbiased features from a biased teacher model. We then transfer these fair features to a student model. We believe that even in biased models, there are some features that are fair and work well for all groups. By focusing on these fair features and avoiding the biased ones, the student model can become both fair and accurate.

The key idea behind MFD is using a special tool called a regularization term. This tool helps to make sure that student features, which depend on different groups, match well with the average features from the teacher model. We measure this match using Maximum Mean Discrepancy (MMD) [1][11]. When the features are aligned like this, it helps keep things consistent within each group and reduces differences between different groups in the feature space. We have two important contributions: First, we introduce the first method of distillation focused on making things fairer [1]. Second, we provide both a deep understanding and practical proof that MFD boosts fairness without lowering accuracy.

II. Related Work

A. Algorithmic Fairness

The field of algorithmic fairness seeks to ensure that automated systems treat individuals equitably, regardless of sensitive attributes. Research in this area has produced a taxonomy of approaches: preprocessing, inprocessing, and postprocessing methods. Pre-processing methods such as fair representation learning [2], Variational Fair Autoencoders [3], and dataset reweighting aim to remove bias from training data. Post-processing methods adjust the model outputs, for instance, by modifying decision thresholds for different groups.

In-processing methods remain the most flexible and theoretically grounded approach. These include adversarial debiasing techniques [4], where an adversary attempts to predict sensitive attributes from model predictions, thereby guiding the main model to unlearn them. Other approaches involve integrating fairness constraints or penalties into the loss function [5][6] or using strategic sampling to ensure balanced representation during training.

However, these approaches require access to raw data or full retraining of the model, making them impractical for systems already deployed or trained on proprietary datasets. In addition, they often introduce a performance-fairness trade-off, where gains in fairness lead to reduced accuracy, highlighting the need for methods that can balance both.

B. Knowledge Distillation and Fairness

Originally introduced by Hinton et al. [7], Knowledge distillation (KD) aims to transfer knowledge from a complex, high-capacity teacher model to a smaller student model for efficient inference. Most KD methods focus on matching output logits (soft labels), attention maps [8], or intermediate feature representations [9]. More advanced statistical methods, such as Neuron Selectivity Transfer [10] and probabilistic knowledge transfer [11], leverage feature distribution alignment using metrics like Kullback-Leibler (KL) divergence or Maximum Mean Discrepancy (MMD).

However, very few of these approaches have considered fairness. Traditional KD methods risk perpetuating and even amplifying biases present in the teacher model, as they aim to replicate behavior without questioning its ethical implications. The recent introduction of MFD addresses this by incorporating fairness into the KD pipeline itself. Our approach avoids copying biased knowledge and instead uses MMD to extract high-utility group-invariant features, producing a student that outperforms baseline fairness methods in both bias reduction and accuracy retention.

III. Models and Datasets

In this study, we experimented with ResNet and ShuffleNet, two well-known deep architectures. ResNet (Residual Network)[17] is a deep architecture known for its skip connections, which help mitigate vanishing gradients and enable the training of very deep networks. It is widely used for image recognition tasks due to its robustness and strong performance. ShuffleNet[18], in contrast, is designed for computational efficiency, employing pointwise group convolutions and channel shuffling to achieve lightweight models suitable for resource-constrained environments.

For the datasets, we decided to go with UTKFace, which differentiates the data based on age, gender, and ethnicity. The UTKFace dataset is a large-scale facial image collection containing over 20,000 images labeled with age, gender, and ethnicity. It covers a broad age range (0–116 years) and includes significant variation in pose, expression, and illumination, making it ideal for studying fairness in facial analysis tasks. Another dataset that we used in our study is CIFAR-10, which is a widely used benchmark for object classification, consisting of 60,000 32x32 color images across 10 distinct classes, such as airplanes, cars, and animals. The dataset is split into 50,000 training and 10,000 test images, providing a standard platform for evaluating model generalization and performance.

IV. Methodology

Fig. 1 shows an outline of our approach to minimize the model size and create a fair model while maintaining comparable accuracy using knowledge distillation. Some of the main questions we need to answer are choosing the fairness criteria and the objective function for the distillation process.

A. Fairness Criterion

There are many fairness criteria, such as overall accuracy equality, fairness through awareness [12], statistical parity [13], equal opportunity, equalized odds, and counterfactual fairness [14]. The choice of the correct fairness criterion depends on the task we are performing and its social, cultural background, since each of them tries to tackle the fairness problem from different aspects.

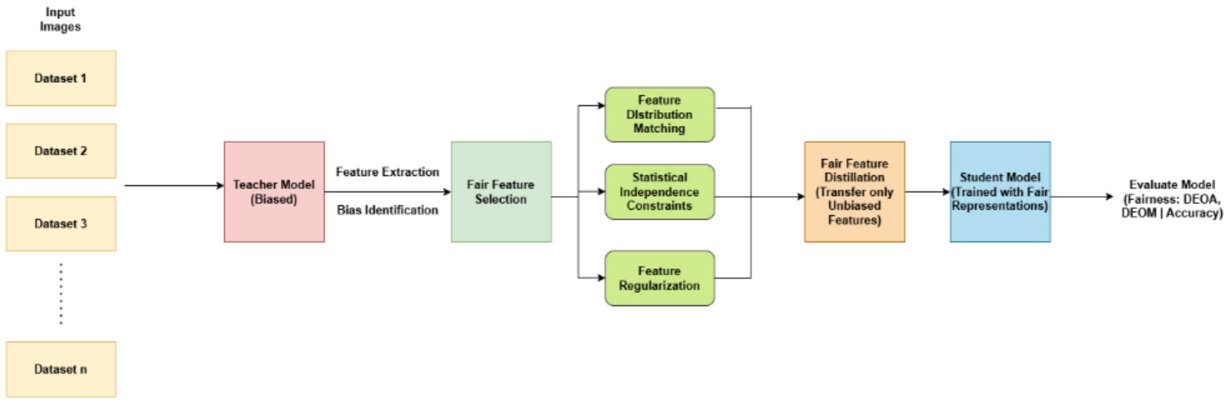
Input Images
Dataset 1
Dataset 2
Dataset 3
Dataset n
Teacher Model (Biased)
Feature Extraction
Bias Identification
Fair Feature Selection
Feature Distribution Matching
Statistical Independence Constraints
Feature Regularization
Fair Feature Distillation (Transfer only Unbiased Features)
Student Model (Trained with Fair Representations)
Evaluate Model (Fairness: DEOA, DEOM | Accuracy)

Fig. 1: Architecture diagram

In this work, we choose the difference in equalized odds (DEO) as the fairness criterion to measure the accuracy discrepancies between groups per class. Equalized odds were initially proposed for the binary case, but can be extended to be a fairness metric by requiring that $\forall a, a' \in \mathcal{A}$, $y \in \mathcal{Y} = \{1, \ldots, M\}$,

$$\Pr(\tilde{Y} = y \mid A = a, Y = y) = \Pr(\tilde{Y} = y \mid A = a', Y = y).$$

. The difference in these probabilities can give us a reliable fairness metric, where a lower value would mean the model is more fair in its predictions. DEO(A) can be formally defined as

$$\mathrm{DEO}_A \triangleq \frac{1}{|\mathcal{Y}|} \sum_y \left(\max_{a,a'} \left(\left| \Pr(\tilde{Y} = y \mid A = a, Y = y) - \Pr(\tilde{Y} = y \mid A = a', Y = y) \right| \right) \right)$$

B. MMD Based Regularizer

We solve the knowledge distillation problem by matching the feature distributions of the teacher and the student model as in [15], rather than minimizing the distances instance-wise since solving the problem from a distributional perspective helps to better solve the fairness problem.

1) Maximum Mean Discrepancy (MMD): For some distributions p and q, MMD is formally defined as follows,

$$\mathrm{D}(\mathbf{p}, \mathbf{q}) \triangleq \sup_{f \in \mathcal{F}} \left(\mathbb{E}_{\mathbf{p}}[f(x)] - \mathbb{E}_{\mathbf{q}}[f(x')] \right) = \|\mu_{\mathbf{p}} - \mu_{\mathbf{q}}\|_{\mathcal{H}},$$

where μ_p and μ_q are the mean embeddings in a reproducing kernel Hilbert space (RKHS). This is a well-defined metric since MMD is 0 only if p = q, that is, the 2 distributions are exactly the same [16].

2) MMD-based Fair Distillation (MFD) Loss: The loss function for MMD-based Fair Distillation (MFD) can be formally defined as

$$\mathcal{L}_{MFD} \triangleq \sum_y \sum_a \mathrm{D}^2(\mathbf{p}_y^T, \mathbf{p}_{a,y}^S),$$

This loss aggregates the squared discrepancy (MMD) between the target distribution (teacher) $\mathbf{p}_y^T$ for class y and the subgroup distributions (student) $\mathbf{p}_{a,y}^S$ y for the sensitive group a and class y.

3) Optimization Function: The final objective function with which we train the student model S is defined as follows

$$\min_\theta \mathcal{L}_{CE}(\theta) + \frac{\lambda}{2} \hat{\mathcal{L}}_{MFD}(\theta),$$

where θ is the model parameter for the student model S. In this equation, $\mathcal{L}_{CE}$ is the regular cross entropy loss of the model and $\hat{\mathcal{L}}_{MFD}$ denotes the MMD-based Fair distillation Loss. λ is an hyperparameter that controls the trade-off between accuracy and fairness seen in the student model. We can apply the MFD loss to several layers of a deep neural network, but our work only focuses on applying it for the feature vectors of the penultimate layer.

V. Expriments & Results

Both ResNet and ShuffleNet were trained and evaluated on these datasets for 20 epochs, enabling a comprehensive assessment of accuracy and fairness trade-offs across different network capacities and dataset characteristics. This setup allows for rigorous comparison of model performance and bias mitigation strategies in visual recognition tasks.

A. *ResNet vs. ShuffleNet on Accuracy and DEO*

#	Model	Model Type	Accuracy	DEO
1	ResNet	Student	73.5%	0.165
2	ResNet	Teacher	72.5%	0.360
3	ShuffleNet	Student	62.0%	0.180
4	ShuffleNet	Teacher	67.5%	0.390

TABLE I: ResNet vs. ShuffleNet

Table 1 presents our experimental results comparing ResNet and ShuffleNet models as both teacher and student networks. We operated these architectures on the UTK Face dataset and the CIFAR-10 for 20 epochs and averaged them.

Our results show that ResNet outperformed ShuffleNet in terms of accuracy, with the student ResNet achieving the highest accuracy at 73.5% compared to ShuffleNet's 62% student model. Also, we observed that while teacher models demonstrated higher DEO (Difference in Equalized Odds) values, student models showed reduced bias with lower DEO scores. ResNet student achieved 0.165 DEO while maintaining higher accuracy than its teacher model, demonstrating effective knowledge transfer with improved fairness. ShuffleNet models showed similar fairness improvements but with more significant accuracy trade-offs. This experiment validates that our distillation approach successfully improves fairness metrics while maintaining competitive accuracy across different network architectures.

B. *ResNet Analysis*

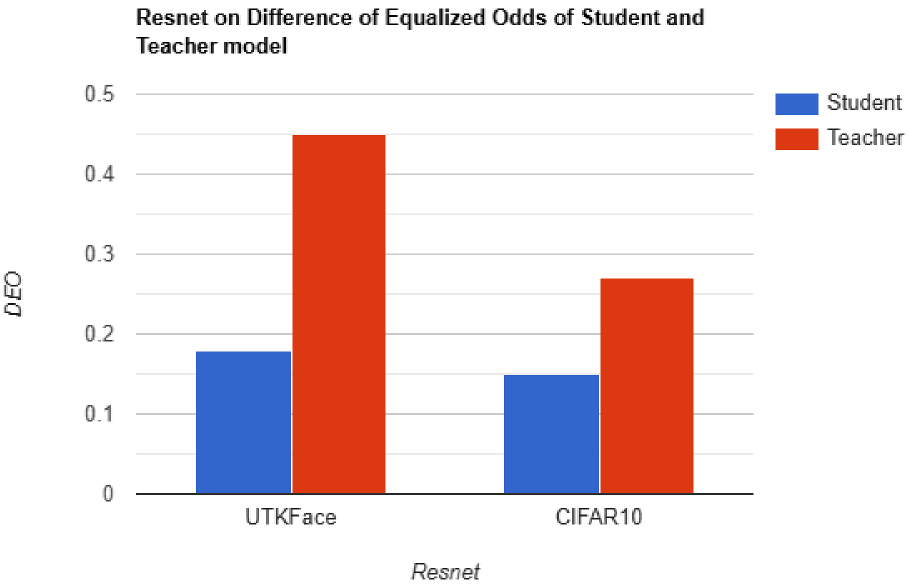

Fig. 2: ResNet Teacher vs. Student: DEO Comparison

Fig. 2. shows that the student ResNet model achieves a substantially lower DEO (0.18 for UTKFace and 0.15 for CIFAR-10) compared to the teacher model (0.45 for UTKFace and 0.27 for CIFAR-10), indicating improved

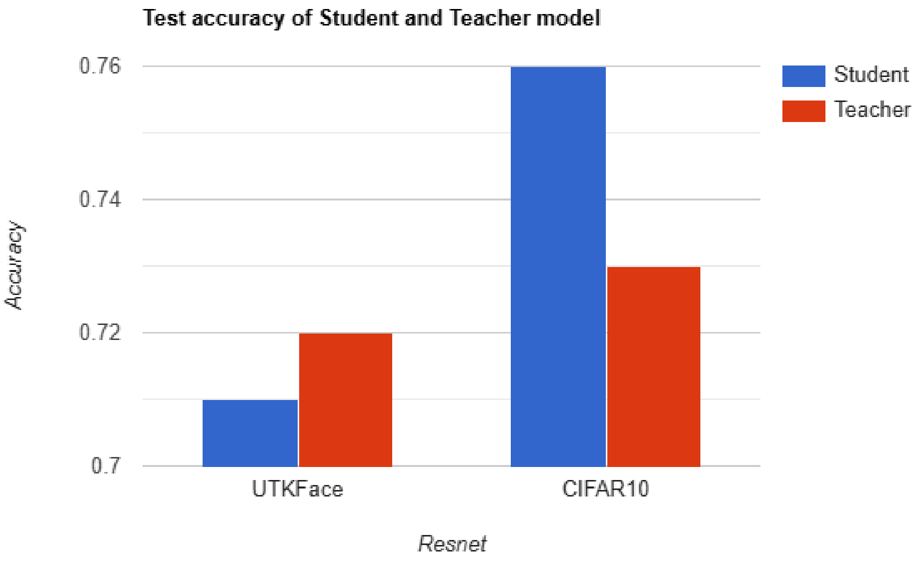

Fig. 3: ResNet Teacher vs. Student: Testing Accuracy Comparison

fairness after distillation. Fig. 3. demonstrates that the student model maintains or even surpasses the teacher's accuracy, especially on CIFAR-10, where the student achieves about 76% accuracy versus the teacher's 73%, while on UTKFace the student is slightly lower (71% vs 72%).

C. ShuffleNet Analysis

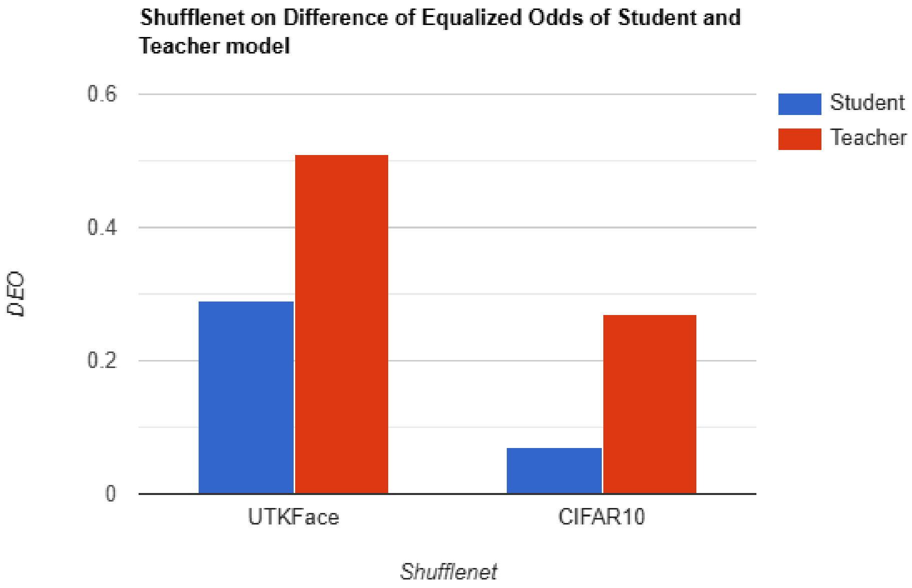

Fig. 4: ShuffleNet Teacher vs. Student: DEO Comparison

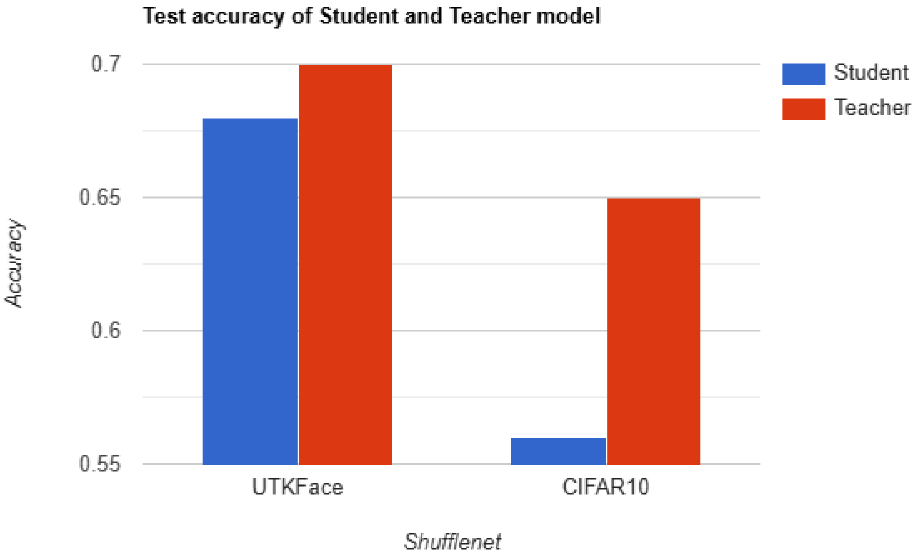

Fig. 5: ShuffleNet Teacher vs. Student: Testing Accuracy Comparison

Fig. 4. demonstrates that the student model achieves substantially improved fairness compared to the teacher model as measured by the difference in equalized odds (DEO). For UTKFace, the student model reduces DEO from approximately 0.5 to 0.28, representing a 44% improvement in fairness. CIFAR-10 shows an even more dramatic fairness improvement, with student DEO (0.07) being nearly 73% lower than teacher (0.27).

Fig. 5. reveals the accuracy trade-offs of this fairness improvement. In UTKFace, the student model maintains competitive accuracy (0.68) compared to the teacher (0.7), with only a minimal reduction of 2. 8%. However, CIFAR-10 shows a more substantial accuracy drop from 0.65 to 0.56 (13.8% decrease) in the student model. This indicates that fairness improvements through distillation maintain better accuracy on UTKFace than on CIFAR-10 when using ShuffleNet architecture.

VI. Discussion

Our work demonstrates that Fair Distillation Based on Maximum Mean Disparities (MFD)[1] is an effective and practical approach to improving fairness in visual recognition models without sacrificing accuracy. Using knowledge distillation, MFD selectively transfers fair and group-invariant features from a biased teacher model to a student model, guided by an MMD-based regularizer. Experimental results in ResNet[17] and ShuffleNet[18] architectures, across synthetic and real-world datasets, show that student models consistently achieve lower Difference in Equalized Odds (DEO) scores - indicating reduced demographic bias - while maintaining or even improving prediction accuracy compared to their teacher counterparts. In particular, ResNet student models achieved the best balance, with significant fairness gains and minimal accuracy trade-offs, while ShuffleNet models also improved fairness but experienced larger accuracy drops. In general, MFD offers a scalable and resource-efficient solution to mitigate bias in AI systems deployed, as it does not require retraining from scratch or access to raw training data. This makes MFD a promising tool for practitioners aiming to deploy fair and reliable computer vision models in high-stakes real-world applications.

VII. Team Contribution

- **Aravinda Jatavallabha** (`arjatava`) — Set up the environment and prepared the UTKFace and CIFAR-10 datasets. Wrote preprocessing scripts and data loaders, trained baseline teacher and student models, performed hyper-parameter tuning, and presented the motivation, model architecture, and training procedure.
- **Aadithya Naresh** (`anaresh`) — Implemented fairness metrics and evaluated model accuracy. Analyzed fairness improvements from teacher to student using MFD, compared ResNet and ShuffleNet on DEO and

test accuracy, produced result tables and plots, trained the initial teacher model, and distilled a smaller student model via Knowledge Distillation.

- **Mihir Arora** (`marora6`) — Implemented the MMD-based Fair Feature Distillation (MFD) technique—including the custom loss function and its integration into the training to balance fairness and accuracy. Authored the MFD motivation, methodology, and theoretical background sections in the presentation and report, and assisted in training both teacher and student models.

References

[1] Jung, S., Lee, J., & Moon, T. (2021). Fair feature distillation for visual recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13274–13283).

[2] Creager, E., Madras, D., Pitassi, T., & Zemel, R. (2019). Fair representation learning by controlling latent space. In Advances in Neural Information Processing Systems (NeurIPS).

[3] Louizos, C., Swersky, K., Li, Y., Welling, M., & Zemel, R. (2015). The variational fair autoencoder. arXiv preprint arXiv:1511.00830.

[4] Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335–340).

[5] Kamiran, F., & Calders, T. (2012). Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), 1–33.

[6] Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In Advances in Neural Information Processing Systems (NeurIPS), 29.

[7] Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

[8] Zagoruyko, S., & Komodakis, N. (2016). Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. arXiv preprint arXiv:1612.03928.

[9] Romero, A., Ballas, N., Kahou, S. E., Chassang, A., Gatta, C., & Bengio, Y. (2014). FitNets: Hints for thin deep nets. arXiv preprint arXiv:1412.6550.

[10] Huang, Z., & Wang, N. (2017). Like what you like: Knowledge distill via neuron selectivity transfer. arXiv preprint arXiv:1707.01219.

[11] Passalis, N., & Tefas, A. (2018). Learning deep representations with probabilistic knowledge transfer. In Proceedings of the European Conference on Computer Vision (ECCV).

[12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, *Fairness through awareness*, In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS)*, 2012, pp. 214–226.

[13] M. Hardt, E. Price, and N. Srebro, *Equality of opportunity in supervised learning*, In *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 29, 2016.

[14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. *Deep learning face attributes in the wild*, In *IEEE International Conference on Computer Vision (ICCV)*, 2015.2, 5,7

[15] Nikolaos Passalis and Anastasios Tefas. *Learning deep representations with probabilistic knowledge transfer.*, In *In European Conference on Computer Vision (ECCV)*, 2018. 2, 3

[16] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Scholkopf, and Alexander Smola. *A kernel two-sample test.*, In *Journal of Machine Learning Research*, 13(1):723–773, 2012. 4

[17] 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 (CVPR), pp. 770-778.

[18] Zhang, X., Zhou, X., Lin, M., & Sun, J. (2018). ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6848-6856.