NSFW-Ninja: Masters of Disguise in the Content Filter Jungle

Neelesh Verma, Jagadeesh Reddy Vanga, Kamalnath Polakam
{neverma, jvanga, kpolakam}@cs.stonybrook.edu

Abstract

The proliferation of social media platforms has led to an increased need for content moderation, with NSFW (Not Safe For Work) detectors playing a pivotal role in maintaining a safe and respectful online environment. However, these detectors are not immune to manipulation, raising concerns about their effectiveness and reliability. This work presents a systematic investigation into the vulnerabilities of NSFW detectors through a black-box attack methodology. Additionally, we present an adversarial attack on the existing NSFW detectors present on popular social media sites. We will systematically perturb the NSFW images and observe the response from the detector. The direction of the perturbation will move the image closer to the detection boundary. In summary, this project seeks to investigate the robustness and expose vulnerabilities in NSFW detectors, demonstrating their susceptibility to manipulation and contributing to the development of more robust and ethical content moderation systems.

1 Introduction

In the digital age, the widespread use of social media platforms has revolutionized the way we communicate, share, and interact with content online. This transformation has brought forth an increased necessity for content moderation, aiming to maintain a safe and respectful online environment for users of all ages. At the forefront of this moderation effort are NSFW (Not Safe For Work) detectors, AI-based systems designed to automatically identify and flag content that may be explicit, offensive, or otherwise unsuitable for public consumption.

These NSFW detectors have become indispensable tools for social media platforms, playing a pivotal role in safeguarding users from inappropriate or harmful content. However, the effectiveness and reliability of these detectors have come under scrutiny in recent years. While they have undoubtedly made significant strides in improving content moderation, they are not impervious to manipulation.

Although NSFW Content consists of text and images, our focus is primarily on images. We will take images that are NSFW originally; and perturb them adversarially such that the final image isn't tagged as NSFW on social media platforms (like Reddit). At the same time, we will also maintain that the perturbation is very small and imperceptible to humans. This is generally done by limiting the L_∞ norm of the perturbation.

Adversarial attacks refer to a class of techniques and methods in the field of machine learning and artificial intelligence (AI) that are used to manipulate or deceive machine learning models. These attacks are designed to exploit vulnerabilities in machine learning algorithms and neural networks in particular. Szegedy et. al. [11] first showed that object recognition systems can be fooled by attacking the MNIST dataset. Since then many different approaches have been developed to exploit detection and recognition systems. In this report, we will assume that the content moderation systems are recognition systems that classify images as NSFW or SFW. Such attacks fall under the white-box and black-box attacks categories. In the white-box scenario, attackers have the ability to access the structures and parameters of the target models, enabling them to create adversarial examples. Conversely, in the black-box scenario, attackers lack any access to the model's structure and parameters. Since we are attacking content moderation systems we cannot access, our method will fall under the black-box setting.

However, existing black-box attack either uses a huge amount of queries [7, 5, 1, 15] or leverages target model training data for a transfer-based attack [31, 29, 16, 30, 9, 27]. We don't have access to the training data of the target model and neither we can perform a huge amount of queries on content moderation systems. Therefore, we opt for a hybrid approach that performs a very small amount of queries and anchors on the transferability of intermediate features across DNN models [20].

2 Literature Review

Adversarial attacks can typically be categorized into two primary categories: white-box attacks and black-box attacks. In a white-box setting, attackers possess detailed information about the victim models, including model structure, parameters, weights, training methodology, etc. In contrast, in a black-box setting, attackers lack access to such information and can only access the output from the model. White-box attacks often rely on exploiting gradient information from the victim model to craft adversarial examples. Prominent white-box attack methods include the Fast Gradient Sign Method (FGSM) [12], Project Gradient Descent (PGD) [18], Carlini and Wagner Attack (C&W) [4], Deepfool [19] and BPDA [2]. However, the white-box attack is unrealistic in our scenario since we don't have access to the content moderation models deployed by the social media sites.

In contrast, black-box attacks occur when attackers do not have access to vital information about the victim model. This setting is more reflective of real-world applications, where the inner workings of the model are hidden from potential attackers. Our attack on content moderation systems also falls under this category. There are mainly two categories of Black-box attacks - *Decision-based attacks* and *Transfer-based attacks*.

2.1 Decision-based attacks

The decision-based attacks query the target model and get the final label (in the classification task) from the target model.

Brendel et al.[3] proposed the first approach that involves a random walk on the decision boundary. In each iteration, the approach randomly selects a direction and projects it onto a boundary sphere to create a high-quality adversarial example, but it's query-intensive and lacks convergence guarantees. Guo et al.[13] proposes Low Frequency Adversarial Perturbation(LFAP) which made few modifications to boundary attack to construct low frequency perturbation. In this method, instead of sampling Gaussian noise low frequency noise is being sampled. By restricting to low-frequency subspace, which has a larger density of adversarial directions, this step succeeds more often, speeding up the convergence towards the target image. Meanwhile, a Query-Limited attack [14] focuses on estimating output probability scores through model queries to transform hard-label attacks into soft-label problems.

On the other hand, [6] takes a different approach, reformulating hard-label attacks as optimization problems aimed at finding the direction that minimizes the distance to the decision boundary. In practical tests, the algorithm efficiently targeted hard-label black-box Convolutional Neural Network (CNN) models on MNIST, CIFAR, and ImageNet, requiring significantly fewer queries (in orders of thousands). The Sign-OPT attack [7] follows a similar optimization approach as [6], treating hard-label attacks as the task of finding the direction with the shortest distance to the decision boundary. Additionally, it efficiently estimated the gradient's sign in any direction, rather than the gradient itself, requiring just a single query. A more recent attack [5] utilized the zeroth-order sign oracle to enhance the Boundary attack, resulting in substantial improvements. They employed a one-point gradient estimate, which, while unbiased, can have higher variance compared to the gradient estimate in [7].

Even though the number of queries required has significantly reduced over the years, their budget is still a big issue due to the extremely small query budget in the content moderation systems. It seems that the robust detection and flagging systems being employed in the most popular social media demand a new attack with just a few queries (in order of a hundred).

2.2 Transfer-based Attacks

Transfer-based black-box attacks are a type of adversarial attack where the attacker generates adversarial examples using a surrogate white-box model and then transfers them to an unknown target black-box model. Transfer-based attacks work because adversarial examples can often fool similar models, and deep learning models are sensitive to small input changes.

Cheng et al.[8] propose a method called Prior-Guided Random Gradient-free(P-RGF) which uses the gradient of a surrogate model as a prior to guide the search direction and then adjusts the direction based on the query feedback from the target model. One of the limitations of the P-RGF method is that it requires a surrogate model that has similar architecture and training data as the target model which might not be possible at times. Another limitation is that it assumes that the target model is deterministic and doesn't have any defense mechanisms such as randomization or gradient masking.

Wang et al.[28] tries to overcome these shortcomings by proposing a novel input transformation-based attack called Structure Invariant Attack(SIA), which applies a random image transformations onto each image block to generate a set of diverse images for gradient calculation. This improves the transferability of the adversarial examples, which can exploit common vulnerabilities of different models and bypass their defenses. Although successful, it may reduce their stealthiness and make them easier to detect by humans, unlike other adversary measures.

A patch-based attack was proposed by Gao et al. [10]. Instead of manipulating the images pixel-wise, it tries adding patches to the image. They incorporated an amplification factor in the FGSM method to increase the step size in each iteration, ensuring that when a pixel's gradient exceeds the ϵ constraint, it is accurately distributed to its neighboring regions through a projection kernel. Zhang et. al. [31] proposed a feature-level attack that employs neuron importance scores. The scores are computed by attributing the model's output to each neuron in the network and total neuron attribution is minimized to craft adversarial examples. But for our scenario, we have no access to the data that is used to train the content moderation models.

All of these transfer-based adversarial attacks assumed that the training data of the target model follows similar distributions as the surrogate model. Zhang et. al. [15] proposed to train the surrogate models in a data-free black-box scenario. They used a GAN for data generation and leveraged model distillation on the substitute model which acts as the discriminator. But they had to make a trade-off with a huge amount of query. For attack on content moderation systems, we have a very small budget for the number of queries (it may be possible to leverage APIs if available to perform such queries but for most of the social media platforms - Reddit, Facebook, Instagram, etc, such APIs aren't available for free). Therefore, we propose a hybrid approach - an extremely limited query-based transfer attack. It has been shown that the DNN models share similar features in their receptive fields [29]. Therefore, even though we don't have the training data, we can generate our own data [21]. The trained model on this data would then share similar features to the target model. We will be leveraging Grad-Cam [22] scores to give scores to the features and the input pixels and minimize this score by querying the target model.

3 Motivation and Scope

The ubiquity of social media platforms has created an urgent need for effective content moderation, particularly in identifying and filtering out NSFW (Not Safe For Work) content. NSFW detectors serve as the first line of defense in protecting users from potentially harmful or inappropriate material. However, their vulnerabilities to manipulation and evasion have raised serious concerns. To our knowledge, there hasn't been any work that tries to perform NSFW-based adversarial attacks. This research is motivated by the imperative to comprehensively understand and address the limitations of NSFW detectors in social media. By conducting black-box attacks on these systems, we aim to uncover vulnerabilities, assess their resilience, and propose strategies for enhancement. Our work seeks to contribute to developing more robust content moderation mechanisms, ultimately fostering safer and more respectful online environments.

4 Implementation Method

To implement our attack, we need a few components. Firstly, we need an NSFW detector so that we can attack it using white-box and then black-box. Moreover, the trained model will provide us with insights into the learned distribution of the NSFW images. Since our dataset is comprised of the images from the NSFW-tagged Reddit images (and SFW images as well), we hope that the trained network will extract transferable features, that can used in the black-box attack. After the trained network, we are doing a white-box attack on our trained model. The reason is that the white-box attack will give us the exploitable features of the images. Although it can be argued that the exploitable features may not be transferable to Reddit, it has been shown that object detectors and image recognition systems look at very similar (or overlapping) image features [32]. One of the primary reasons is the use of transfer learning in different vision tasks.

Our pipeline can be broadly divided into 3 parts - i) Training an NSFW detector, ii) Performing a White-box attack on the trained detector, and iii) Transfer the attack to Reddit. We will be going through each of these 3 stages in detail in the following sections.

4.1 Training an NSFW Detector

4.1.1 Network Architecture

The first task is to train an NSFW detector. There are many architectures to select from - MobileNets, EfficientNets, ResNets, InceptionNets, VGGs, etc. We looked for architectures that have been currently deployed in content moderation systems. We found out that Bumble [23] has released a private detector for detecting lewd images on Bumble chat. However, it's trained on a different dataset and has a slightly different task at hand (it's not exactly an NSFW detector but rather for lewd images that are more common in Bumble chat). But this gave us a starting point. They have used EfficientNet-v2 [25] as their base architecture. We are instead using pre-trained Inception-Resnet-v2 (trained on ImageNet) [24] as it is a better choice where generalization to new images is critical. Since the diversity of NSFW images varies a lot, we want a more generalized model, and Inception-Resnet-v2 fits into that category. We replaced the final layer (top layer) of the Inception-Resnet-v2 with 2-dense layers. This way, these 2 layers will learn NSFW features and we can exploit these learned features for future tasks (transferring to black-box attack). The network architecture for the training is shown in the figure 1.

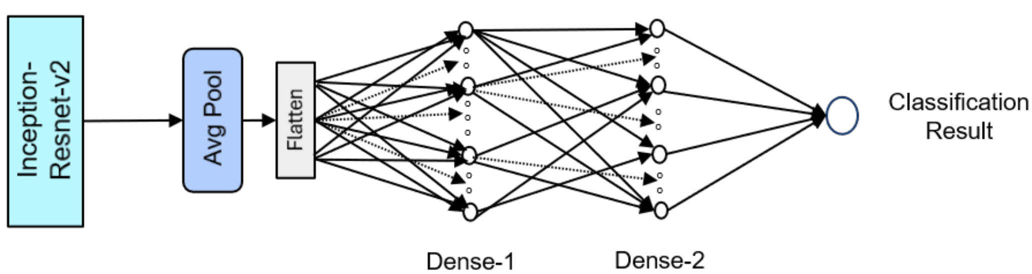

Figure 1: Training Pipeline for NSFW detector

First, the image is passed through a pre-trained Inception-Resnet-v2 model (with the top layer removed). Then, the output is fed into two dense layers (Dense-1 and Dense-2) via an average pooling in between. At the end, we take the outputs from the Dense-2 layer and pass them through the final dense layer to get a single output. We take the sigmoid of the output to get a probability value which we consider as an NSFW score (0 means SFW and 1 means NSFW).

4.1.2 Dataset

One of the primary challenges of our task is to make a dataset. There isn't any publicly available dataset of Reddit NSFW images. We tried generating NSFW images [21]. The problem with such methods was that the generated images were not always tagged as NSFW by Reddit. It means that there is a separate definition of NSFW on Reddit as compared to these images generated. To get the actual images that are tagged as NSFW by Reddit, we made a web-scraper that will crawl through different NSFW subreddits and scrape the images. Although this method should work in theory, there are a few issues with it - Images may get deleted by the moderators, and the posted images are not NSFW at all. It is easy to solve the first issue in which Reddit posts a template image whenever an image is deleted.

The second issue is more complicated - images are not NSFW at all. Since we have nearly 0.1 million images, it is not feasible for us to go through the images manually. Instead, we used a mixture of multiple methods - Bumble private detector [23], a Python-based NSFW detector [1], and Google's SafeSearch [2]. Even though none of these detect NSFW (according to Reddit), using a combination of these we were able to filter out images that will be considered NSFW whatsoever (extremely explicit images). This way, we constructed our dataset - 52200 training images, and 5800 testing images.

4.1.3 Training and Evaluation of NSFW Detector

We divided the training set into - 46400 training and 5800 validation images. The training epochs are 1000, optimizer is Adam-W [17] with weight decay as 0.01, learning rate 0.001 and β_1, β_2 as 0.9 and 0.99 (used as standard). The images are resized to 256×256, and the loss function is binary cross-entropy. The Inception-Resnet-v2 is pre-trained on ImageNet and Dense-1 layer has 256 neurons and the Dense-2 layer has 128 neurons with ReLU activations in each of them.

The training accuracy for 1000 epochs is 92.7 % and the test accuracy came out to be 88.1 %. Although both accuracies can be further improved, we are not interested in making a perfectly accurate NSFW detector but rather exploiting it.

4.2 White-box attack on the trained detector

Our attack is based on the heat map generated by the Grad-CAM. Grad-CAM generates a heat map that quantifies the importance of different regions/pixels in an image that contributed to the classification of that particular image. Our intuition was that we could add noise gradually to the important region (as described by Grad-CAM) of an NSFW image and ultimately decrease the scores of these pixels. Another approach based on entropy [26] has been employed earlier for detecting attacks. So the noise will be added to the image with weights described by the Grad-CAM heat-map.

Assume that the image is given by x and the corresponding heat-map generated by Grad-CAM as $h(x)$. Also, assume that our NSFW detector is f. We have $f(x) = 1$. We will add noise $n(x)$ iteratively such that -

$$\text{minimize}\|n(x)\|_2 \quad \text{s.t.} \quad f(x + \alpha h(x) n(h(x))) = 0 \tag{1}$$

So, we are minimizing the noise matrix (such that it remains imperceptible to humans) such that this noise added to image x weighted by the heatmap $h(x)$ is misclassified. Here α is a regularization factor. We have chosen n to be Gaussian with learnable parameters - μ and Σ. This way, we can iteratively add noise and update the parameters accordingly. $n(h(x))$ will always add more noise in the heat maps with red regions (higher values) if we choose μ to be the centroid of the red region (after thresholding the heat map). The covariance matrix is determined by the spread of the red region (in x- and y-direction).

For now, we are getting decent results (if not very good) from our attack. Since our optimization is per-image-based, we were not able to run it on all the images. Till now, we were able to run it on 1000 NSFW test images and were able to successfully attack 751 of these.

[1] https://github.com/GantMan/nsfw_model

[2] https://cloud.google.com/vision/docs/detecting-safe-search

5 Current State and Obstacles

Currently, we are in the stage of completing our White-box attack (almost completed, only evaluation is done on a small subset yet). We haven't begun transferring the attack to Reddit although we have begun some preparations. For now, we are adding adversarial patches to these NSFW images at different locations and posting them on Reddit. By observing the behavior of Reddit on a single image using various adversarial patches at different locations, we are trying to learn if there are certain regions in the image that the Reddit content moderation is paying specific attention to. We are sure that such regions do exist as shown by Grad-CAM [22]. If we can find such regions in a couple of images, we can then use those regions for the attack (these regions will serve as kind of a heatmap of equation 1).

The obstacle is only to post various such images on Reddit to make any good observations. It is a bit time-taking and manual procedure (we might have to post 100 images) but we should be able to do it.

6 Evaluation Plan

The final evaluation plan will be to take an NSFW image, add noise to it that is imperceptible to humans, and then post it to Reddit. It should not be tagged as NSFW for a successful attack. We are planning to produce thousands of such images.

7 Timeline

The approximate timeline is as follows -

- Complete the evaluation of the White-box attack — Nov 15, 2023
- Post NSFW images on Reddit with patches attached — Nov 20, 2023
- Final black-box optimization based on the learning from the above step — Dec 1, 2023
- Final report and presentation — Dec 5, 2023

References

[1] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search, 2020.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

[3] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[5] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1277–1294. IEEE, 2020.

[6] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018.

[7] Minhao Cheng, Simranjit Singh, Patrick Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2019.

[8] Shuyu Cheng, Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Improving black-box adversarial attacks with a transfer-based prior. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019.

[10] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*, pages 307–322. Springer, 2020.

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[12] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2014.

[13] Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. In *Uncertainty in Artificial Intelligence*, pages 1127–1137. PMLR, 2020.

[14] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *International conference on machine learning*, pages 2137–2146. PMLR, 2018.

[15] Jianghe Xu Shuang Wu Shouhong Ding Lei Zhang Chao Wu Jie Zhang, Bo Li. Towards efficient data free black-box adversarial attack. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15115–15125, 2022.

[16] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

[18] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

[19] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.

[20] Muzammal Naseer, Salman H Khan, Shafin Rahman, and Fatih Porikli. Task-generalizable adversarial attack based on perceptual metric. *arXiv preprint arXiv:1811.09020*, 2018.

[21] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022.

[22] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[23] Semen Zhydenko Stepher O' Farrell, Gleb Vazhenin. Bumble releases open-source version of private detector a.i. feature to help tech community combat cyberflashing. 2022.

[24] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

[25] Mingxing Tan and Quoc Le. Efficientnetv2: Smaller models and faster training. In *International conference on machine learning*, pages 10096–10106. PMLR, 2021.

[26] Bilel Tarchoun, Anouar Ben Khalifa, Mohamed Ali Mahjoub, Nael Abu-Ghazaleh, and Ihsen Alouani. Jedi: Entropy-based localization and removal of adversarial patches. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4087–4095, 2023.

[27] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021.

[28] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4607–4619, 2023.

[29] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1158–1167, 2020.

[30] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019.

[31] Jianping Zhang, Weibin Wu, Jen-tse Huang, Yizhan Huang, Wenxuan Wang, Yuxin Su, and Michael R Lyu. Improving adversarial transferability via neuron attribution-based attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14993–15002, 2022.

[32] Barret Zoph, Vijay Vasudevan, Jonathon Shlens, and Quoc V Le. Learning transferable architectures for scalable image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8697–8710, 2018.