Toward Universal Detector for Synthesized Images by Estimating Generative AI Models

Ryota Seo*, Minoru Kuribayashi†, Akinobu Ura*, Antoine Mallet‡, Rémi Cogranne‡,
Wojciech Mazurczyk§ David Megías¶
* Okayama University, Japan, E-mail: {pr0765pj, pl843doc}@s.okayama-u.ac.jp
† Tohoku University, Japan, E-mail: kminoru@tohoku.ac.jp
‡ Troyes University of Technology, France, E-mail: {antoine.mallet, remi.cogranne}@utt.fr
§ Warsaw University of Technology, Poland, E-mail: wojciech.mazurczyk@pw.edu.pl
¶ Internet Interdisciplinary Institute (IN3), Universitat Oberta de Catalunya, Spain, E-mail: dmegias@uoc.edu

***Abstract*—One of the vulnerabilities in discriminators for AI-generated images is that the classification accuracy degrades when dealing with images generated using methods other than those they were trained on. As a countermeasure, in this study, we propose an image generation method estimator. The process of discrimination involves the input of an image to the estimator, which estimates the method used for its generation. Subsequently, a specialized fake image discriminator tailored to the estimated image generation method is used to identify the authenticity of the image. The activation functions are also considered according to the estimation results and analyzed for those discriminators. Discrimination scores are weighted and aggregated according to the estimation results, and the final decision is output. Our experimental results showed that the estimator achieved a classification accuracy of approximately 90% for 18 types of AI-generated images. Furthermore, by selecting the top two estimations in order of confidence, the accuracy increased to around 98%.**

I. Introduction

Since the introduction of Generative Adversarial Networks (GANs) in 2014 [1], numerous image generators have been developed [2]–[6], generating images so sophisticated that they are indistinguishable from those created by humans. Unlike the GAN-based models, diffusion models such as Stable Diffusion and Midjourney use forward and reverse processes to generate a composite image [7].

As a forensic technique, it has been studied to assess whether a given content is naturally captured by a camera or artificially produced. At the early stage of image generators, some artifacts are observed by human eyes if we carefully observe the synthesized images. With the development of machine learning techniques, such obvious artifacts no longer appear. CNN-based detectors have been studied as alternatives to the human eye, and quantitative and qualitative evaluations have been conducted on various types of datasets.

However, it raises a new challenge for the practical application of forensic techniques for analyzing artifacts, such that discrepancies between training and test data lead to poor detector performance. In the case of a deepfake detector, several detection methods achieve high test accuracy on one dataset, but show low detection accuracy on datasets that differ from the training phase. If the detector recognizes artifacts outside the training data set, the performance in terms of classification accuracy significantly degrades. It is thus necessary to evaluate the generalization of the classifiers.

A weakness of the currently available fake image detectors is that their classification accuracy drops significantly for the images generated using methods different from those encountered during their training. Mismatches in the training and testing phases are considerably difficult problems when discriminating against fake images. On the other hand, if a CNN-based detector is trained with various images generated by several generation methods, the classification accuracy becomes lower.

This study proposes a new universal fake image detection framework that achieves high classification accuracy, even with input generated by a method different from the training datasets. One of the advantages is the low computational cost of retraining new image generation methods that may emerge in the future. We introduce the idea of an "Estimator" that classifies an input image into groups according to how they were generated. The proposed framework consists of two main steps. First, an estimate is made of the input image on how the image was generated. Then, the image is classified as fake or real using each discriminator specialized for the corresponding generation method.

The discriminator dedicated to identifying specific still image generators is not limited to only the method with the highest confidence output by the Estimator but utilizes multiple discriminators based on activation output. This approach compensates for any misestimations made by the Estimator at the discrimination step. For such activation, three different algorithms are considered in this study. The first method is "Static," which focuses on each value of the confidence vector output by the Estimator. The second method, "Top-β," focuses on the number of special discriminators used for each input image. The third method is "Dynamic," which focuses on the sum of the values in the confidence vector output by the Estimator for the image generation method used for activation.

In the experiments, we design the estimator using a fine-tuned model of XceptionNet [8] pre-trained on ImageNet [9]. We validate the proposed estimator's ability to infer the generation method from a given image by employing 18 image generation methods, including GAN [1] models such as

StyleGAN [6] and GigaGAN [10], as well as diffusion models [11] such as Stable Diffusion [12] and DALL-E [13].

II. Related Works

In this section, we provide an overview of image generation techniques relevant to this study and review the difficulty of the forensics study.

A. Generative Adversarial Networks

Generative Adversarial Networks (GANs) [1] involve two networks: the Generator, which generates images from random noise, and the Discriminator, which distinguishes between generated images and original images. The Generator aims to generate images that can deceive the Discriminator into classifying them as genuine images rather than fake ones. On the other hand, the Discriminator strives to minimize a loss function that enables it to accurately classify the images generated by the Generator as fake images.

1) StyleGAN: In StyleGAN [6], the structure of the Generator is modified. In traditional GANs, the Generator produces images from random noise. In StyleGAN, the input layer is omitted, and generation starts from pre-trained constants. Noise images are supplied to each synthesis network's layer. After each convolution in the Generator, normalization is performed using Adaptive Instance Normalization (AdaIN) [14]. Similar to PGGAN, StyleGAN also utilizes Progressive Growing for training.

StyleGAN has undergone various improvements over time. To eliminate the droplet-like noise observed in generated images, StyleGAN2 [15] was introduced, which utilizes weight demodulation with standard deviation-based normalization instead of AdaIN [14]. Lastly, there is StyleGAN3 [16], which ensures that all layers of the Generator are equivariant with respect to continuous signals, intending to generate images even more natural than those produced by StyleGAN2.

B. Latent Diffusion Models

A diffusion model [11] defines a diffusion process by simply adding easily manageable noise to input data and learning the data distribution by reversing the diffusion process over a finite time. Diffusion models using denoising autoencoders can generate highly sophisticated images. However, optimizing such a powerful diffusion model requires an enormous amount of time and resources. Therefore, a method called Latent Diffusion Models (LDM) was proposed in [17] to achieve learning of diffusion models under limited computational resources. Numerous improved models of LDMs [18]–[21] have been published, including Stable Diffusion [12], which reduces computational time and resources by separating the compression learning and generative learning phases. Other models include DALL-E 2 [13], which combines the multimodal model CLIP [22] with the diffusion model GLIDE [23].

C. Mismatch at Training and Testing

The study in [24] examined the effectiveness of various feature extraction algorithms in four automated facial manipulation techniques (Deepfakes, Face2Face, FaceSwap, NaturalTextures) [25]. High recognition accuracy is demonstrated when the training and testing methods are identical. However, when different methods are used for training and testing, the recognition accuracy decreases significantly.

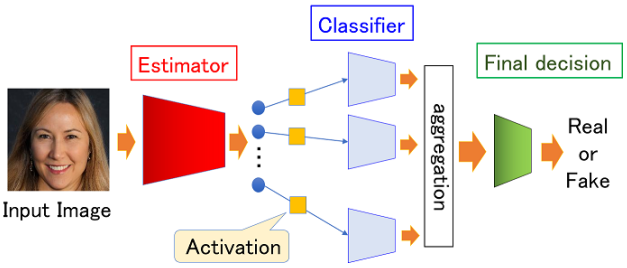

Fig. 1. Overview of proposed detector composed of the estimator of generation method and classifier.

Furthermore, the approach in [26] involves using a classifier trained using an enhanced spectrum and an RGB image-trained detector for assessment. When trained using images generated by the Progressive Growing of GANs (PGGAN), this method exhibits high recognition accuracy across multiple models with network structures similar to PGGAN. Data augmentation is introduced in [27], [28] to overcome dataset bias. However, the empirically designed augmentation strategies have certain limitations due to the poor generalizations.

III. Proposed Method

In this section, we propose an architecture for assessing images generated by image-generating AI systems. It comprises an *Estimator* and a *Classifier*. In the proposed method, we employ an estimation model $\mathrm{Est}(\cdot)$ and a set of discriminator models $(D_1(\cdot), D_2(\cdot) \ldots, D_n(\cdot))$, and the overall process is illustrated in Fig.1.

The Estimator assesses the method used to generate the input image. The estimated generation method corresponds to a fake image discriminator that evaluates whether the image is real or fake. In the classifier process, the estimated generation methods are not limited to the one with the highest output probability; instead, multiple top-ranked methods are selected and provided to the discriminator for assessment.

A. Estimator of Generation Method

An input image I is assessed using an estimation model $\mathrm{Est}(\cdot)$ to determine which generation method was applied to produce it. The model is trained using images generated by n different generation methods and outputs a confidence vector $\boldsymbol{p}$ of length n, where the i-th element is the probability that the input image was generated by the i-th generation method:

$$\boldsymbol{p} = (p_1, \ldots, p_n) = \mathrm{Est}(I), \tag{1}$$

where $\sum p_i = 1$.

In this study, we employ fine-tuning on XceptionNet, which has been pre-trained on ImageNet, to generate the estimation model $\mathrm{Est}(\cdot)$. In fine-tuning, the fully connected layer is removed and a newly designed fully connected layer is added along with a dropout layer and dense connections to output the confidence vector $\boldsymbol{p}$. As the default input image size for

TABLE I
HYPER-PARAMETERS IN THE PROPOSED ACTIVATION FUNCTIONS.

Method	Hyper-parameter
Static	T
Top-β	$\beta \in \mathbb{Z}_n$
Dynamic	$0 < \gamma \leq 1$

XceptionNet is 299×299, respectively, an input image I is resized to adjust these dimensions before being processed. It is worth mentioning that the fine-tuning models employed in this study can be replaced with a sophisticated model to improve performance.

When the input image is not produced by a generation method but is actually a captured photograph, the expected outcome for the confidence outputs is random. However, it does not affect the final decision due to the use of the classifier mentioned in the following section.

B. Classifier with Some Discriminators

First, we apply an activation function $\mathrm{Act}(\cdot)$ to the confidence vector $\boldsymbol{p}$. It selects the discriminator $D_i(\cdot)$ corresponding to each image generation method based on $\boldsymbol{p}$. According to a threshold T, $(0 < T \leq 1)$ determined in $\mathrm{Act}(\cdot)$, we calculate the binary vector $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n) = \mathrm{Act}(\boldsymbol{p}) \tag{2}$$

where

$$\alpha_i = \begin{cases} 1 & \text{if } p_i \geq T \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Hence, the selection is not limited to one specific discriminator, but a group of discriminators. The details of the activation function are explained in the following section.

Next, for $1 \leq i \leq n$, the binary classification result $b_i \in \{0, 1\}$ is calculated by using the i-th discriminator $D_i(\cdot)$:

$$b_i = \alpha_i D_i(I) \tag{4}$$

Note that $b_i = 1$ implies that I is judged as fake by the discriminator $D_i(\cdot)$. If the input image is actually a real photograph, the result is expected to be $b_i = 0$ for all discriminators, hence, the "error" at the Estimator does not matter during the final decision.

At the final decision, the final probability P is calculated by aggregating the binary results $\boldsymbol{b}$ with the confidence vector $\boldsymbol{p}$ as follows:

$$P = \frac{\sum_{i=1}^{n} b_i p_i}{\sum_{i=1}^{n} \alpha_i p_i}. \tag{5}$$

If $P > 0.5$, the input image I is judged fake; otherwise real.

It is worth mentioning that each discriminator $D_i(\cdot)$ can be selected from conventional works, considering both the classification accuracy and computational complexity.

C. Activation

In this study, the following three methods for setting the threshold T are devised and compared. Table I enumerates the hyper-parameters utilized in these methods.

TABLE II
ACCURACY OF ESTIMATOR'S TOP-RANKED CONFIDENCE LEVELS.

	Estimator
Top1 Accuracy	0.900
Top2 Accuracy	0.982
Top3 Accuracy	0.994

1) Static Method: It is a simple method that employs a pre-determined threshold T without considering the confidence vector $\boldsymbol{p}$.

2) Top-β Method: In this method, the elements with the top β confidence values p_i are activated. Here β is a pre-determined parameter. At first, the confidence vector $\boldsymbol{p}$ is sorted in descending order:

$$\bar{\boldsymbol{p}} = (\bar{p}_1, \ldots, \bar{p}_n) = \mathrm{sort}(\boldsymbol{p}), \tag{6}$$

where $\mathrm{sort}(\cdot)$ denotes sorting the elements in descending order. Then, the threshold is selected as the β-th elements from $\bar{\boldsymbol{p}}$.

$$T = \bar{p}_\beta \tag{7}$$

3) Dynamic Method: Different from the Top-β technique, the dynamic method determines β according to the sum of their confidence values.

First, we calculate the sorted confidence vector $\bar{\boldsymbol{p}}$. Then, the sum of top confidence values is calculated, and β is determined for a pre-determined parameter γ:

$$\beta = \arg\min_i \left(\sum_i \bar{p}_i \geq \gamma \right) \tag{8}$$

Finally, the threshold T is selected using Eq.(7).

IV. EXPERIMENTS

In our experimental evaluation, we measure the classification accuracy of the estimator designed by fine-tuning XceptionNet. We used datasets consisting of images generated by 18 different image generation methods and real-world images.

We conducted two experiments. First, we investigated the accuracy of the Estimator. Second, we examined the overall accuracy of the proposed method in the presence of unknown image generators.

A. Estimation of Image Generation Method

In this experiment, we used 24-bit RGB color images generated by $n = 18$ different image generation methods. The resolution of generated images can be 512×512, 768×768, or 1024×1024 pixels, which are larger than the input size of XceptionNet (299×299 pixels). Therefore, each image was cut into 9 blocks of 299×299 pixels chosen from the top left, top center, top right, middle left, middle center, middle right, bottom left, bottom center, and bottom right of the image. We use 63,000 cropped blocks from 7,000 images for training, while 18,000 blocks from 2,000 images are used for testing. The estimated generation method was determined by averaging the confidence scores from the 9 cropped blocks.

Figure 2 describes a confusion matrix for identifying image generation methods. It arranges actual generation method on

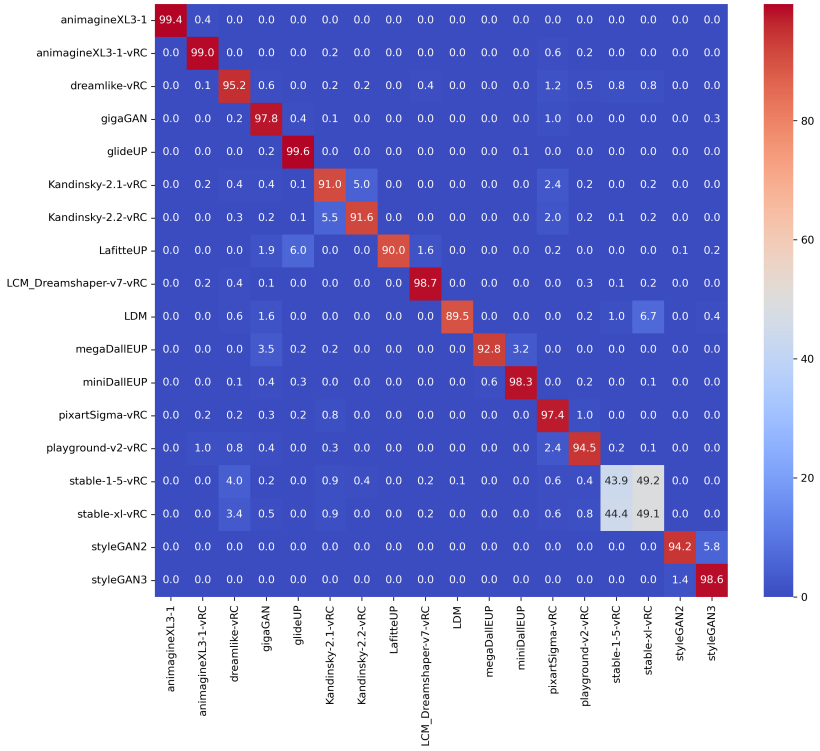

Fig. 2. Confusion matrix for estimator

the vertical axis and estimated generators on the horizontal axis, showing the accuracy with which the Estimator correctly classified image generation method output by each label ($2,000$ images per label). The classification accuracy was approximately 90%. In the case of two versions of stable diffusion method, it occurs misclassification due to the similar characteristics appeared in the generated images. In the second step of the proposed method, such misclassification is not a problem because the corresponding discriminator $D_i(\cdot)$ can output binary classification results with high accuracy. Additionally, as shown in Table II, the "Top2 Accuracy" was around 98%, indicating that the estimator accurately classified the image generation method within the Top2 class. Based on the above results, the Estimator demonstrates high classification accuracy, allowing the proposed method to achieve high accuracy comparable to using all classifiers specialized for each image generation method, while incurring lower computational costs.

B. Comparison of Activation Methods

If an image generated by an unknown image generation method, the confidence vector $\boldsymbol{p}$ calculated by Estimator is expected to vary across n classes and each discriminator $D_i(\cdot)$ may not accurately classify whether it is generated. To simulate such a situation, we train the Estimator excluding one specified image generation method and measure the performance of the proposed method for the images generated by the excluded image generation method.

In this experiment, we compared three activation methods

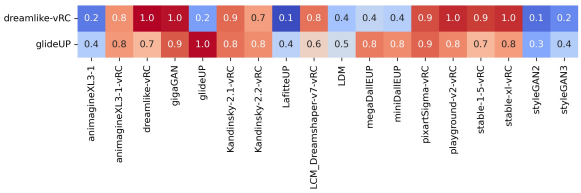

Fig. 3. Comparison of confidence value as outputs of two discriminators "Dreamlike" and "Glide".

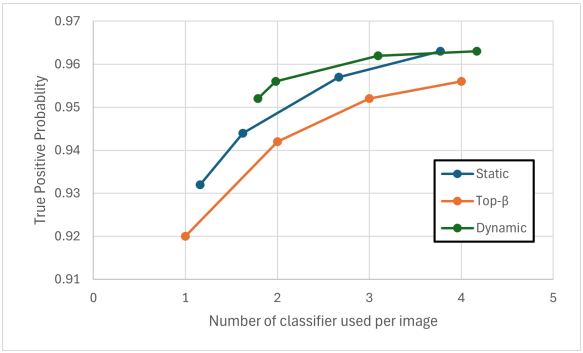

Fig. 4. Comparison of three activation methods in terms of true positive probability by using images generated by "Glide".

for images generated by 18 different image generation methods and real photographs. We train each discriminator $D_i(\cdot)$, which is a fine-tuned version of the XceptionNet, using the images generated by each image generation method. These

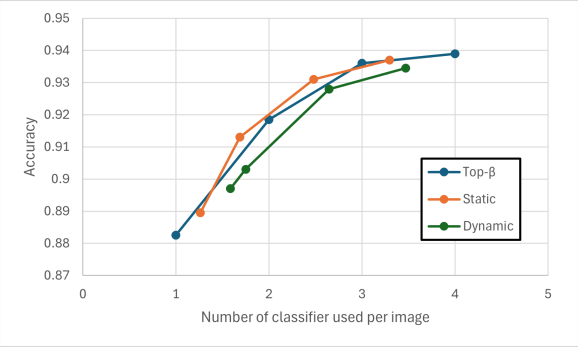

Fig. 5. Comparison of three activation methods in terms of classification accuracy in the case of "Glide".

discriminators were designed to perform binary classification between images generated by each image generation method and real photographs from the FFHQ dataset [29]. For training, we used 63,000 cropped images of size 299×299 pixels for each label, using the same method as the Estimator. Note that, when activation was not used, the accuracy was approximately 99%.

First, as a preliminary experiment, we investigated how confident each discriminator, specialized in a particular image generation method, was when classifying images generated by other methods. The investigation was conducted using 2000 images for each method. Due to the page limitation, we only show the results of two image generation methods "Dreamlike[1]" and "Glide[2]" in Fig. 3. It is observed that the classification accuracy is not necessarily low for images generated by unknown image generation methods. A discriminator specialized for one image generation method can be said to be compatible to a certain extent with other image generation methods in terms of classification accuracy.

In both cases, the Estimator is trained by excluding a specific image generation method and outputs $\boldsymbol{p}$ with $n = 17$. In the case of "Dreamlike," some mismatched discriminators show similar classification accuracy to the matched discriminators. Therefore, the overall accuracy is close to the matched (Dreamlike) case, and there was no difference in accuracy due to the activation methods.

On the other hand, in the case of "Glide," the classification accuracy of some top discriminators tends to be lower than that of the matched discriminator. Therefore, the selection of a good $D_i(\cdot)$ plays an important role, which is controlled by $\mathrm{Act}(\boldsymbol{p})$. The true positive probability using 1000 images generated by "Glide" is shown in Fig. 4, and the classification accuracy using 1000 real images and 1000 images generated by "Glide" is shown in Fig. 5. For images generated by "Glide," the Top-β method specifies a fixed number of $D_i(\cdot)$, thus ignoring the information extracted by the Estimator. While the Static method does not consider the choice of threshold T, the Dynamic method provides flexibility in selecting the threshold. This improves the stability of predictions and the appropriateness of the number of activated discriminators in the proposed method. On the other hand, for real photographs, it is found that the Dynamic method needs to activate relatively more classifiers compared to the other two methods.

Experimental results show that when a particular discriminator, such as "Dreamlike", exhibits high classification accuracy, there is little need for Activation. This is because the misclassification correction capability provided by Activation is well complemented by other discriminators and only results in additional computational costs. On the other hand, when the classification accuracy of some of the top discriminators is lower than that of the matched discriminators, as seen in "Glide", Activation has been confirmed to be effective in improving classification accuracy and can be used in such a situation.

Based on the results, the Static method is considered the most suitable among the three proposed methods for both real photographs and generated images. However, as the Dynamic method shows better accuracy for generated images, developing more flexible algorithms is necessary. In particular, it is important to optimize the balance between computational efficiency and accuracy by more effectively utilizing the confidence information obtained from the Estimator to improve adaptability to unknown image generation methods. Additionally, the real photograph data used may be biased, so further experiments with more diverse, raw data are required to improve the reproducibility of the results.

V. Conclusions

In this paper, we proposed a two-stage classification architecture for detecting synthesized images created by a generative AI model. An estimator was introduced to improve the classification accuracy of a detector for synthesized images. We envision a workflow that estimates several possible image generation methods from a suspicious image and aggregates the binary classification results obtained from each discriminator to determine whether the image is a fake image or not.

In the experiment, we designed an estimator using a model that is a fine-tuned version of the XceptionNet. The use of other models and tuning the parameters to further improve accuracy is a future task. In this experiment, the estimator was also classified by image generation method. However, in the future, by integrating image generation methods with similar features, the estimator aims to reduce the number of classes to classify and at the same time improve the accuracy.

Acknowledgment

The authors acknowledge the funding obtained from the EIG CONCERT-Japan call to the project Detection of fake newS on SocIal MedIa pLAtfoRms "DISSIMILAR" through grants PCI2020-120689-2 (Agencia Estatal de Investigación, Spain), EIG CONCERT-JAPAN/05/2021 (National Centre for Research and Development, Poland), and JPMJSC20C3 (Japan Science and Technology Agency, Japan).

The work of D. Megías was supported, in part, by the "SECURING" project (PID2021-125962OB-C31) funded by

[1]https://huggingface.co/dreamlike-art/dreamlike-diffusion-1.0
[2]https://github.com/openai/glide-text2im

the Ministry of Science and Innovation, the Agencia Estatal de Investigación, and the European Regional Development Fund (ERDF), and by the ARTEMISA International Chair of Cybersecurity (C057/23) and the DANGER Strategic Project of Cybersecurity (C062/23), both funded by the Spanish National Institute of Cybersecurity through the European Union – NextGenerationEU and the Recovery, Transformation, and Resilience Plan. The work of M. Kuribayashi was supported by the JSPS KAKENHI (22K19777).

References

[1] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[2] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *CoRR*, 2017. arXiv: 1710.10196.

[3] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. ICCV'17*, 2017, pp. 2242–2251.

[4] Y. Choi, M. Choi, M. Kim, J. Ha, S. Kim, and J. Choo, "StarGAN: Unified generative adversarial networks for multi-domain image-to-image translation," in *Proc. CVPR'18*, 2018, pp. 8789–8797.

[5] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. ICLR'19*, 2019.

[6] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," *IEEE Trans. Pattern Analysis & Machine Intelligence*, vol. 43, no. 12, pp. 4217–4228, 2021.

[7] C. Zhang, C. Zhang, M. Zhang, and I. S. Kweon, "Text-to-image diffusion models in generative ai: A survey," *CoRR*, 2023. arXiv: 2303.07909.

[8] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR'17*, 2017, pp. 1800–1807.

[9] J. Deng, W. Dong, R. Socher, K. L. L. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. CVPR'09*, 2009, pp. 248–255.

[10] M. Kang, J.-Y. Zhu, R. Zhang, *et al.*, "Scaling up gans for text-to-image synthesis," in *Proc. CVPR'23*, 2023, pp. 10 124–10 134.

[11] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. ICML'15*, vol. 37, 2015, pp. 2256–2265.

[12] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," *CoRR*, 2021. arXiv: 2112.10752.

[13] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *CoRR*, 2022. arXiv: 2204.06125.

[14] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. ICCV'17*, 2017, pp. 1510–1519.

[15] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR'20*, 2020, pp. 8107–8116.

[16] T. Karras, M. Aittala, S. Laine, *et al.*, "Alias-free generative adversarial networks," in *Proc. NeurIPS'21*, 2021.

[17] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. CVPR'22*, 2022, pp. 10 674–10 685.

[18] Y. Zhou, R. Zhang, C. Chen, *et al.*, "Towards language-free training for text-to-image generation," in *Proc. CVPR'22*, 2022, pp. 17 907–17 917.

[19] J. Chen, J. Yu, C. Ge, *et al.*, "Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis," *CoRR*, 2023. arXiv: 2310.00426.

[20] J. Chen, C. Ge, E. Xie, *et al.*, "Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation," *CoRR*, 2024. arXiv: 2403.04692.

[21] A. Razzhigaev, A. Shakhmatov, A. Maltseva, *et al.*, "Kandinsky: An improved text-to-image synthesis with image prior and latent diffusion," *CoRR*, 2023. arXiv: 2310.03502.

[22] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. ICML'21*, 2021, pp. 8748–8763.

[23] A. Nichol, P. Dhariwal, A. Ramesh, *et al.*, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *CoRR*, 2021. arXiv: 2112.10741.

[24] Y. Xu and S. Y. Yayilgan, "When handcrafted features and deep features meet mismatched training and test sets for deepfake detection," *CoRR*, 2022. arXiv: 2209.13289.

[25] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "Faceforensics++: Learning to detect manipulated facial images," in *Proc. ICCV'19*, 2019, pp. 1–11.

[26] M. Tanaka, S. Shiota, and H. Kiya, "A universal detector of CNN-generated images using properties of checkerboard artifacts in the frequency domain," *CoRR*, 2021. arXiv: 2108.01892.

[27] L. Li, J. Bao, T. Zhang, *et al.*, "Face X-ray for more general face forgery detection," in *Proc. CVPR'20*, 2020, pp. 5000–5009.

[28] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, "Learning self-consistency for deepfake detection," in *Proc. ICCV'21*, 2021, pp. 15 003–15 013.

[29] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR'19*, 2019, pp. 4401–4410.