Source Attribution for Images Generated by Diffusion-Based Text-to-Image Models: Exploring the Forensics Approach

Xinqi Jiang[1], Lingyun Tian[1], Teng Sio Hong[2], Jinyu Tian[1*]
[1] Macau University of Science and Technology, Macau
E-mail: {3240003534, 2230015805}@student.must.edu.mo, jytian@must.edu.mo
[2] Faculty of Humanities and Social Sciences, Macao Polytechnic University, Macau
E-mail: p1308598@mpu.edu.mo

***Abstract*—As the image generation technology continues to advance, images generated by artificial intelligence have become ubiquitous on the internet, leading to a plethora of controversies regarding image copyright. Addressing this issue, tracing back to the source model of generated images has become a crucial approach to resolving it, aiding in establishing the legitimacy and ownership of images. Most existing approaches primarily focus on discerning image authenticity while overlooking the issue of source attribution for generated images. Few traditional image attribution methods are constrained by temporal limitations, only applicable to attributing past generative models such as GANs, while their effectiveness is limited for newer methods like diffusion generative models that have emerged in recent years. This paper proposes a novel method capable of attributing images generated by text-to-image diffusion models and maintaining effectiveness even for untrained models. We construct a semantic feature extraction network and train a feature-difference recognition model using a siamese network approach based on the semantic similarity of generated images. We collect and construct a dataset for training and testing, validating the outstanding performance of our method. Our method enhances compatibility with black-box models, ensuring effective source identification for images generated through text input. Through rigorous experimental validation, our method demonstrates significant progress, providing effective forensics approach for attributing generated images.**

I. Introduction

Over recent years, significant advancements have been observed in text-to-image generation models. Notably, models such as Stable Diffusion [1], DALL-E [2], Imagen [3], and others have demonstrated the capability to produce high-quality, lifelike images. A model in this domain receives a prompt, characterized as a textual segment, along with random noise as inputs. Subsequently, the model executes denoising of the image, guided by the provided prompt, ensuring the resultant image aligns with the given prompt.

With the proliferation of generated images flood the internet, there arises controversy concerning copyright issues and the difficulty in tracing their origins [4]. Some advertisers have used unauthorized models to generate commercial advertising images, and the authors of the models hope to receive payment from the advertisers. However, proving that these images were generated by their models is an extremely challenging task, as attributing the exact source of image generation can be restricted by technical limitations and data privacy [5]. The example highlights the complex challenges posed by copyright disputes and traceability issues in an era of AI-generated images. Consequently, research on attributing synthetic image sources is crucial for forensic investigations aimed at addressing infringement activities. Additionally, it plays a pivotal role in fostering the establishment of a more equitable and intellectual property-respecting society.

However, although the considerable work [6]–[10] devoted to identifying fake images, research on tracking remains limited [11]–[13], with the majority of existing studies focused on images generated by GAN [14]-based models. Most method for detecting images generated by GAN-based model exhibit limited efficacy when applied to diffusion-based generation models [15], rendering them inadequate for traceability.

In this paper, our main contributions are as follows:

- We propose a novel method for source attribution of synthetic images generated by diffusion-based text-to-image models, which can be naturally applied to real attribution scenarios.
- We propose a supervised contrastive loss that can effectively decouple the semantic features and style features of images.
- We constructed a strongly semantically related dataset, and our model performed well on the test dataset.

II. Related Work

A. Text-to-Image Generation

Text-to-image generation typically involves taking a text description (prompts) as input and producing an image that corresponds to the text. Early pioneering efforts in text-to-image generation [16] utilized Generative Adversarial Networks (GANs) [14]. These approaches combined a prompt embedding with a latent vector, aiming for the GANs to generate an image that illustrates the prompt. Such early work [17]–[19] has inspired further research in text-to-image generation using GANs. However, GAN-based models do not always deliver optimal performance in image generation [1]. More recently, significant advancements in text-to-image generation have been achieved with the introduction of diffusion models [1]–[3],

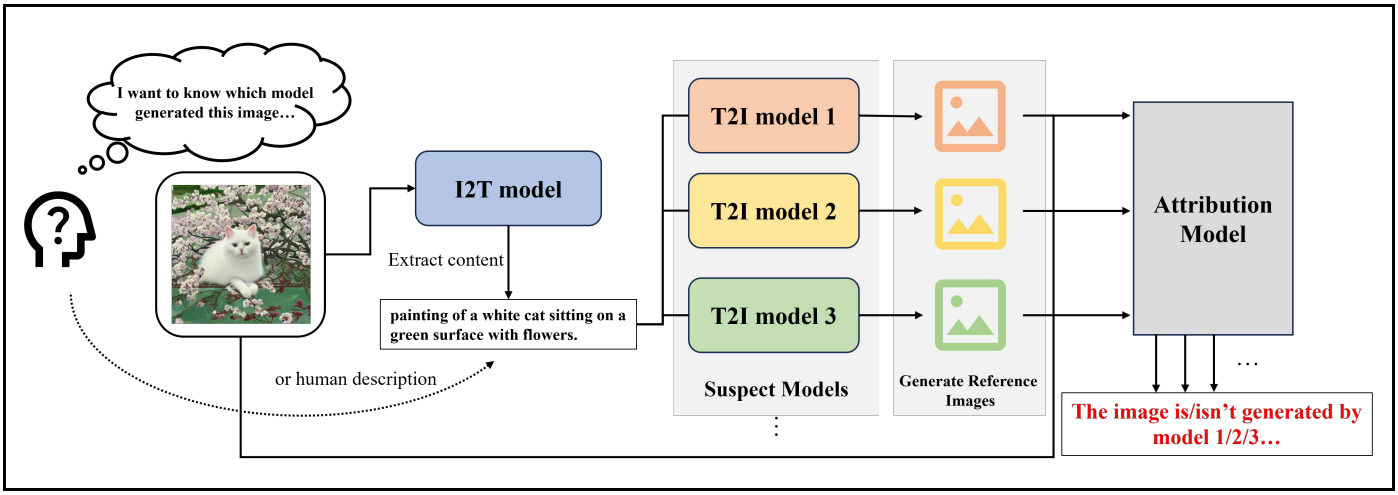


Fig. 1: Overview of Image Source Attribution Process: User first needs to describe the image to be attributed or automatically obtain descriptions using the I2T model. The obtained description is then input into the suspicion model to generate a reference image. These two images are then input together into the attribution model to output result.
I2T model: Image to Text model. **I2T model**: Text to Image model.

[20], [21]. These models typically start with random noise and prompts, then iteratively refine the noisy images into clear ones guided by the prompts. Modern diffusion model-based approaches, such as Stable Diffusion [1], DALL·E [2], Imagen [3], and GLIDE [21], have set new benchmarks in performance, surpassing earlier models. Consequently, our work focuses on these diffusion-based models.

B. Synthetic Image Attribution

Over the past year, research on the traceability of fake images has mainly focused on tracing images generated by GAN-based models. Representative works include: [11] propose a GAN fingerprinting technique based on representation mixing, which can matching images invariant to their semantic content. Their solution demonstrates robustness against benign transformations (e.g., changes in quality, resolution, shape) typically encountered during online image resharing. [12] present an iterative algorithm for discovering images generated from previously unseen GANs by exploiting the fact that all GANs leave distinct fingerprints on their generated images. [13] also uses the unique model fingerprint existing in the GAN, and will leave a stable fingerprint in the generated image, supporting the image attribution. Their experiments show that even minor differences in GAN training can result in different fingerprints, which enables fine-grained model authentication.

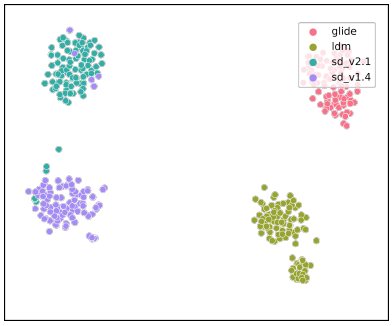
However, the results of these methods on the emerging T2I models are disappointing [15]. The literature concerning the tracing of Text-to-Image (T2I) generation models remains sparse, with only a handful of studies addressing this area. Notably, among the limited existing literature, the works of [15] and [22] stand out as significant contributions.

Sinitsa et al. [15] utilize the inductive bias of convolutional neural networks (CNNs) to develop a new detection method that requires a small amount of training samples and achieves accuracy that is on par or better than current state-of-theart methods.Based on the observation that images produced by each generative model exhibit distinct model fingerprints [13], [15] aims for the network to internalize this unique characteristic. Unlike analogous methodologies, the author utilizes a network to generate this Residual. Specifically, throughout model training, a similarity computation is conducted between the artifacts produced by the network and both synthetic and authentic images from the same source. The objective is to minimize the disparity between the artifacts and synthetic images while maximizing the gap between the artifacts and authentic images. Consequently, a novel loss function is introduced to facilitate this objective. During the inference phase, assessing the similarity solely between the acquired artifacts and the target image enables decision-making. Nonetheless, this approach encounters a notable challenge—its vulnerability to noise, as confirmed by the conducted experiments. Adding even a minimal amount of noise to the image is adequate to compromise the model's effectiveness.

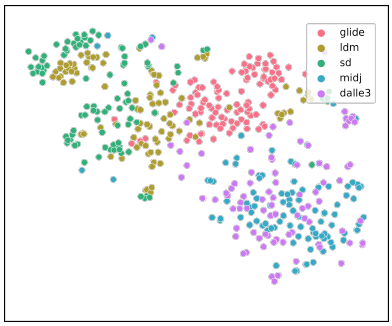
III. Method

A. Motivation

Our approach is based on the following hypothesis : A Text-to-Image (T2I) model is expected to exhibit a certain level of semantic and style similarity when generating images for the same prompt. Specifically, when employing the T2I model to generate images based on identical prompts, the resultant images demonstrate a notable correlation in content semantics and visual style, such as painting style, color brightness, etc. Furthermore, distinctions in content and visual semantics among various models arise, due to disparities in image databases and training methodologies employed during model



(a) Generated images from the same prompt but different models.



(b) Random prompts and different models.

Fig. 2: Utilize the image encoder of the CLIP model for extracting features from image data, then apply the t-SNE method to dimensionality reduction and visualization.

training. This unique and distinguishable similarity forms the basis for establishing a feature discrimination network, which aims to learn the semantic differences among images generated by different models under the same prompt.

To validate this hypothesis, we employed the Image Encoder of the CLIP [23] model for verification. Initially, we utilized a same prompt to generate 100 images for each of the four different T2I models (GLIDE [21], LDM [1], SDv1.4 [24], SDv2.1 [25]). Part of the sample image are shown in Fig.3. Subsequently, these images were input into the Image Encoder to obtain corresponding feature vectors. Finally, we employed the t-SNE [26] method for dimensionality reduction and visualization, as illustrated in Fig.2a. It is evident that, even with the use of non-finetuned pretrained weights, there is still a significant distinction among the models. Notably, the SDv1.4 and SDv2.1 models exhibit substantial separability, with SDv2.1 being fine-tuned on the basis of SDv1.4, demonstrating that their respective generated image features remain well-differentiated.

On the other hand, for each of the 5 different generative models (GLIDE, LDM, SDv2.1, Midjourney, DALLE·3), we randomly sampled 100 images from publicly generated datasets. Repeating the aforementioned procedure, the obtained results are depicted in Fig.2b. It is evident that their distributions are quite scattered and challenging to differentiate.

B. Process Design

Based on the above experimental observations, an apparent approach is to utilize identical prompts to generate multiple fake images, subsequently clustering them for tracing. However, this method is impractical in real-world scenarios due to the substantial time consumption involved in generating the images. The time cost required for image tracing through clustering is unacceptable. Therefore, we propose a feasible solution for real-world scenarios to tracing the origin of images generated by T2I models. When a user intends to trace the origin of a particular image, they can initially obtain the textual content of the image through self-description or an Image-to-Text (I2T) model (such as Blip [27], gpt2 [28] ...). Subsequently, this textual description is fed into the Text-to-Image(T2I) model under consideration for detection. The resulting newly generated image is then compared with the original image using a discriminant model, ultimately yielding a conclusion. If the image is generated by the T2I model under consideration, the newly generated image should exhibit strong semantic similarity to the original image. In such cases, the discrimination network would provide an affirmative decision. Conversely, if neither of them holds true, the discrimination network would render a negative decision for each pair. Thus, we convert the task of source attribution to a binary classification problem. The overall process of our approach is shown in Fig.1.

We believe that contrasting the input image with the newly generated image is more persuasive and practical than using a model to make direct judgments. After all, innocent until proven guilty.

C. Model Design

The above experiments have verified that the image encoder of CLIP possesses strong semantic feature extraction capabilities, and the provided features have separability at different generated model levels. Therefore, the detection model can be constructed by a feature extractor and a classification network. Meanwhile, considering our goal is to assess the similarity of two images in terms of content and visual semantics, we adopt the idea of Siamese networks [29]. The input images are separately processed by the encoder to obtain two features, and the distance between them is calculated to obtain the model loss. The specific model design is illustrated in Fig.4.

Here, we propose a Supervised Contrastive Loss [30], designed to decouple the semantic features and style features of images, defined as follows:

$$\mathcal{L}_{SupCon} = \frac{1}{|I|}\sum_{i\in I}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(d(\mathcal{F}_i,\mathcal{F}_p)/\tau)}{\sum_{a\in A(i)}\exp(d(\mathcal{F}_i,\mathcal{F}_a)/\tau)}, \tag{1}$$

where I denote the set of images generated using the same prompt, $P(i)$ the set of all images sharing the same source as i, $A(i)$ the set of all images from different model than i, $A, P \in I$. F_i denote the features of the image indexed by i, τ the scaling factor, and $d(,)$ represent the difference between two features.

Fig. 3: Image samples generated by different models from the same prompt: *'In a garden, a white cat rests beneath cherry blossoms, sunlight casting shadows amidst swirling petals'.*

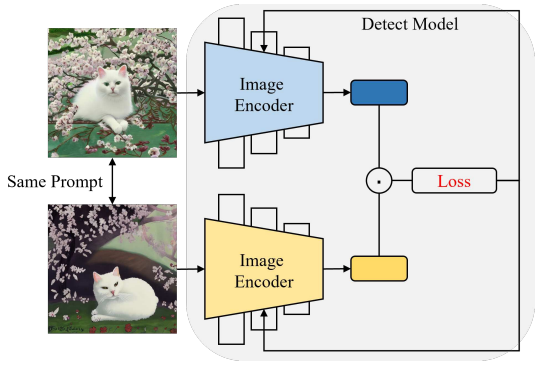


Fig. 4: Structure of the detect model.

TABLE I: Structure of dataset.

Type	Image Source	Image Size	Training Data	Testing Data
Real	Flickr	Not fixed	6742	1349
Fake	LDM	256 × 256	6742×3	1349×3
	SDv2.1	768 × 768	6742×3	1349×3
	SDv1.4	768 × 768	6742×3	1349×3
Total	-	-	67420	13490

IV. Experiment

A. Data Collection

We chose the Flickr [31] dataset as the real image training data. On the other hand, we utilize the generated images from LDM, SDv14 and SDv2.1 models for training. Specifically, we employ the LDM, SDv14 and SDv2.1 models to generate images corresponding to the text from the Flickr dataset. The final training dataset is shown in Table.I.

Here 6742×3 indicates that, for the same prompt, the generative model produced three synthetic images. The total number of images in our training data is 67,420, and testing data is 13,490.

B. Implementation Details

We decompose the attribution task into a binary classification problem, where given a pair of images as input, the model must determine whether the two images are generated by the same generative model. We use CLIP [23] and HR-Net [32] as backbones to extract image features. Each model was trained for 100 epochs using the Adam optimizer, with an initial learning rate set to 1e-4. The learning rate was scheduled to decay by a factor of 0.4 every 20 epochs to ensure stable convergence. The experiments were carried out on a single Nvidia A6000 GPU, leveraging the PyTorch framework for implementation.

C. Result

Table.II presents our experimental results. From the table, it is evident that CLIP and HR-Net without fine-tuning are ineffective; they tend to classify all image pairs as originating from the same source. In contrast, after fine-tuning CLIP and HR-Net using our method, the accuracy of provenance detection improves. This can be explained by the fact that images with the same prompt have strong semantic similarity, which CLIP and HR-Net can capture these similarities to make the same-source judgment. Our method further enables the model to focus more on the style features of the images, thus decoupling the original features. Notably, the provenance results with HR-Net as the backbone are significantly better than with CLIP. This is because the CLIP model aims to align image content with text, making it difficult to separate style and semantic features, whereas HR-Net focuses more on the image itself, thus better capturing the desired style features.

Fig.5 shows the probability distribution of predictions for the test data using HR-Net. It can be observed that, except for the image pairs consisting of real and sdv21 models, our

TABLE II: Results of different backbones on the test sets.

Backbone	Model1	Model2	TP	FP	FN	TN	Recall	F1 Score	Accuracy	AUC
CLIP w/o fine-tune	LDM	SDv14	4842	7263	0	0	1	0.57	40.00%	0.82
	LDM	SDv21	4842	7263	0	0	1	0.57	40.00%	0.78
	SDv14	SDv21	4842	7263	0	0	1	0.57	40.00%	0.59
	Real	LDM	2421	2421	0	0	1	0.57	50.00%	0.89
	Real	SDv21	2421	2421	0	0	1	0.57	50.00%	0.87
	Real	SDv14	2421	2421	0	0	1	0.57	50.00%	0.83
CLIP	LDM	SDv14	3807	1414	1035	5849	0.78	0.75	72.92%	0.86
	LDM	SDv21	3833	1345	1009	5918	0.79	0.76	74.02%	0.86
	SDv14	SDv21	3832	5363	1010	1900	0.79	0.54	41.67%	0.54
	Real	LDM	1904	323	517	2098	0.79	0.82	85.50%	0.81
	Real	SDv21	1929	1467	492	954	0.80	0.66	56.80%	0.63
	Real	SDv14	1903	1324	518	1097	0.79	0.67	58.97%	0.66
HR-Net w/o fine-tune	LDM	SDv14	4842	7263	0	0	1	0.57	40.00%	0.79
	LDM	SDv21	4842	7263	0	0	1	0.57	40.00%	0.82
	SDv14	SDv21	4842	7263	0	0	1	0.57	40.00%	0.59
	Real	LDM	2421	2421	0	0	1	0.67	50.00%	0.88
	Real	SDv21	2421	2421	0	0	1	0.67	50.00%	0.88
	Real	SDv14	2421	2421	0	0	1	0.67	50.00%	0.78
HR-Net	LDM	SDv14	4576	37	266	7226	0.95	0.98	99.20%	0.99
	LDM	SDv21	4541	6	301	7257	0.94	0.97	99.87%	0.99
	SDv14	SDv21	4307	869	535	6394	0.89	0.86	83.21%	0.95
	Real	LDM	2405	15	16	2406	0.99	0.99	99.38%	0.99
	Real	SDv21	2136	557	285	1864	0.885	0.83	79.32%	0.90
	Real	SDv14	2171	244	250	2177	0.90	0.90	89.90%	0.96

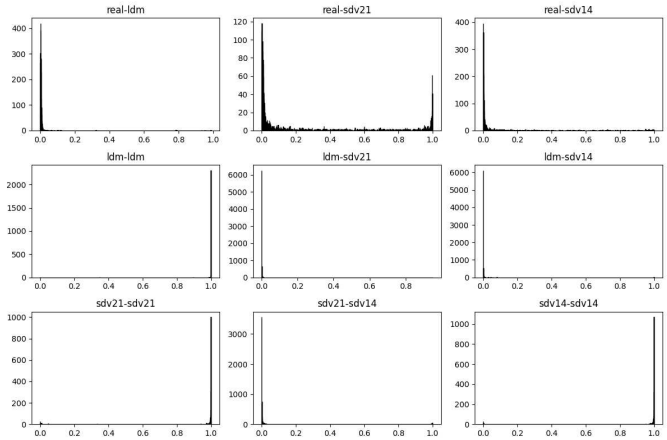


Fig. 5: Probability distribution of HR-Net backbone predictions on test data.

method demonstrates high stability in other cases. For the real and sdv21 image pairs, we attribute this to the fact that the sdv21 model's training data is larger compared to the other generative models, resulting in generated images whose style more closely approximates real images, making them harder

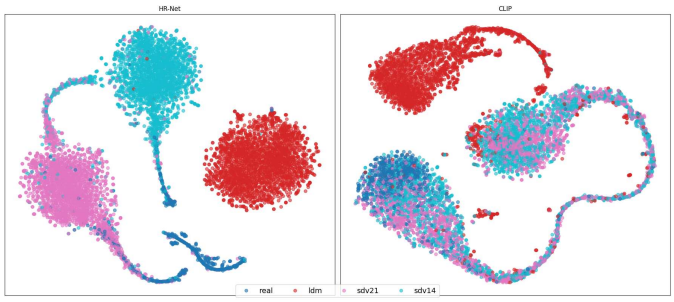


Fig. 6: T-SNE visualization of test image features extracted by CLIP and HR-Net.

to distinguish from real images.

In Fig.6, we visualize the image features extracted by the two backbones using t-SNE. It is evident that HR-Net distinguishes the features of different models, with only few features of real images overlapping. In contrast, CLIP performs rather poorly.

V. Conclusion

In this paper, We presents a novel approach for source attribution of generated images produced by diffusion-based

text-to-image models. Compared to existing methods, our approach is more friendly towards black-box models. Users only need to input the same image descriptions into these models and then compare the generated images with the original ones. After experimental testing, our model exhibited strong performance in image source attribution task. This provides a means of proof for copyright maintenance among image content creators.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[2] J. Betker, G. Goh, L. Jing, *et al.*, "Improving image generation with better captions," *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, vol. 2, no. 3, p. 8, 2023.

[3] C. Saharia, W. Chan, S. Saxena, *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in neural information processing systems*, vol. 35, pp. 36 479–36 494, 2022.

[4] C. Campbell, K. Plangger, S. Sands, and J. Kietzmann, "Preparing for an era of deepfakes and ai-generated ads: A framework for understanding responses to manipulated advertising," *Journal of Advertising*, vol. 51, no. 1, pp. 22–38, 2022.

[5] S. Chesterman, "Good models borrow, great models steal: Intellectual property rights and generative ai," *Policy and Society*, puae006, 2024.

[6] H. Song, S. Huang, Y. Dong, and W.-W. Tu, "Robustness and generalizability of deepfake detection: A study with diffusion models," *arXiv preprint arXiv:2309.02218*, 2023.

[7] J. Lu, Y. Li, J. Zhou, B. Li, and S. Lyu, "Forensics forest: Multi-scale hierarchical cascade forest for detecting gan-generated faces," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2023, pp. 2309–2314.

[8] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "Cnn-generated images are surprisingly easy to spot... for now," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 8695–8704.

[9] J. Zhang, Y. Wang, H. R. Tohidypour, and P. Nasiopoulos, "Detecting stable diffusion generated images using frequency artifacts: A case study on disney-style art," in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 1845–1849.

[10] K. Shiohara and T. Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

[11] T. Bui, N. Yu, and J. Collomosse, "Repmix: Representation mixing for robust attribution of synthesized images," in *European Conference on Computer Vision*, Springer, 2022, pp. 146–163.

[12] S. Girish, S. Suri, S. S. Rambhatla, and A. Shrivastava, "Towards discovery and attribution of open-world gan generated images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 14 094–14 103.

[13] N. Yu, L. S. Davis, and M. Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7556–7566.

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[15] S. Sinitsa and O. Fried, "Deep image fingerprint: Accurate and low budget synthetic image detector," *arXiv preprint arXiv:2303.10762*, 2023.

[16] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*, PMLR, 2016, pp. 1060–1069.

[17] N. Bodla, G. Hua, and R. Chellappa, "Semi-supervised fusedgan for conditional image generation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 669–683.

[18] Z. Wang, Z. Quan, Z.-J. Wang, X. Hu, and Y. Chen, "Text to image synthesis with bidirectional generative adversarial network," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, 2020, pp. 1–6.

[19] D. M. Souza, J. Wehrmann, and D. D. Ruiz, "Efficient neural architecture for text-to-image synthesis," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–8.

[20] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

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

[22] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "De-fake: Detection and attribution of fake images generated by text-to-image generation models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, 2023, pp. 3418–3432.

[23] A. Radford, J. W. Kim, C. Hallacy, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, PMLR, 2021, pp. 8748–8763.

[24] *Stable diffusion v1-4 model.* https://huggingface.co/CompVis/stable-diffusion-v1-4.

[25] *Stable diffusion v2-1 model.* https://huggingface.co/stabilityai/stable-diffusion-2-1.

[26] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[27] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, PMLR, 2022, pp. 12 888–12 900.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[29] G. Koch, R. Zemel, R. Salakhutdinov, *et al.*, "Siamese neural networks for one-shot image recognition," in *ICML deep learning workshop*, Lille, vol. 2, 2015.

[30] P. Khosla, P. Teterwak, C. Wang, *et al.*, "Supervised contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 18 661–18 673, 2020.

[31] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Transactions of the Association for Computational Linguistics*, vol. 2, pp. 67–78, 2014.

[32] J. Wang, K. Sun, T. Cheng, *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.