Latent Landscapes: Topology of the CLIP Feature Space for Synthetic Image Detection

1st Lea Uhlenbrock
FAU Erlangen-Nürnberg
Erlangen, Germany
lea.uhlenbrock@fau.de

2nd Sandra Bergmann
FAU Erlangen-Nürnberg
Erlangen, Germany
sandra.daniela.bergmann@fau.de

3rd Christian Riess
FAU Erlangen-Nürnberg
Erlangen, Germany
christian.riess@fau.de

***Abstract*—The rapid advancement of AI-generated images poses a growing challenge to image authenticity in critical domains such as journalism and law enforcement. Current detection methods struggle with unseen generative models and post-processing manipulations, making robust detection increasingly important. In this work, we investigate the robustness properties of the feature space of Contrastive Language-Image Pretraining (CLIP) for synthetic image detection. To this end, we analyze the topological structure of synthetic and real images within CLIP's latent space and study how post-processing attacks influence their geometry. Our findings reveal that the positioning of training datasets might give clues to their suitability for generalization and that synthetic and real images react differently to specific manipulations, creating distinguishable features. Further, certain transformations shift samples toward a fixed point in feature space, creating a certain level of predictability of post-processing shift. By investigating these effects, we provide new insights into the CLIP feature space and its role in improving the generalizability and robustness of AI-generated image detection.**

***Index Terms*—synthetic images, detection, deepfake, CLIP, robustness**

I. Introduction

AI-generated images have become a part in digital culture, influencing entertainment, marketing, and political messaging. As tools have become more advanced and accessible, even non-experts can create flawless synthetic images. With the vanishing of visual flaws and the rapid progress of AI image generation, image authenticity is threatened in fields where it is crucial: journalism, law enforcement, insurance fraud or fake news detection. Research on synthetic image detection has intensified, with many methods targeting sources like GANs, autoencoders, and diffusion models. However, most detectors struggle with unseen sources and post-processing. In real-world settings, where images are often compressed or altered, robust detection remains critical. A promising direction is classifying synthetic images using the feature space from Contrastive Language-Image Pretraining (CLIP) [1], which has shown strong performance and robustness to low-level perturbations, while generalizing better to unseen images than traditional methods [1]–[5]. However, to the best of our knowledge, a thorough analysis of the latent properties of synthetic and real images in regard to robustness properties remains missing.

In this work, we explore the topological properties of synthetic and real images within CLIP's feature space and analyze how post-processing attacks influence latent geometry, mapping out factors that may influence classification performance: Certain operations, such as blurring or noise, shift features toward a fixed point, while others move samples closer to the decision boundary, increasing the likelihood of misclassification. We also observe that synthetic and real images respond differently to certain manipulations. Building on this insight, we design a compact set of features based on post-processing behavior to improve generalization performance. From our explorative investigation, we hope to provide first steps towards understanding and cartographing the CLIP space more thoroughly for the robust detection of synthetic images.

Work was supported by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as part of the Research and Training Group 2475 "Cybercrime and Forensic Computing" (grant number 393541319/GRK2475/2-2024).

II. Related work

Research interest in the detection of synthetic images has been rising rapidly, which lead to the discovery of many different forensic traces. The following literature review is necessarily limited in scope, and we refer to recent surveys for a broader coverage of the topic [6].

Statistical methods are arguably the most popular approach to the detection of synthetic images. These methods are based on the fact that image generators have until now failed to fully synthesize the statistics of natural images. Internally, they oftentimes calculate pixel statistics to expose generated images [7]–[14]. However, these traces are rather fragile and can be highly individual per image source. Deep representations of images based on large pre-trained networks can be used to extract traces that generalize well across generator architectures. One such line of methods relies on CLIP [1]. CLIP is trained on an extensive corpus of text-image pairs, leading to semantically meaningful embeddings of the scene.

Several works have used this representation for synthetic image detection. Ojha *et al.* [2], Lin *et al.* [3], and Khan *et al.* [5] report the advantages of CLIP representations for distuingishing real and generated images. Moskowitz *et al.* [15] and Cioni *et al.* [16] leverage CLIP-features for source attribution of synthetic images. CLIP can also be used for inpainting detection [17]. Furthermore, Cozzolino *et al.* [4] show that

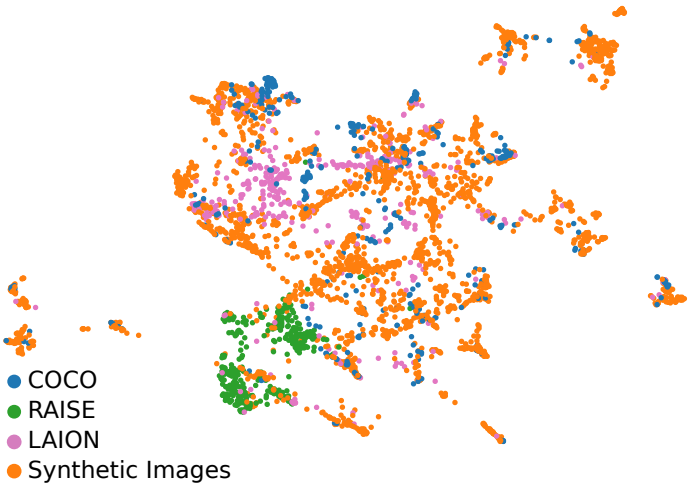


Fig. 1: Distribution of synthetic and real samples in CLIP feature space, projected onto 2D using UMAP.

even a small sample size and lightweight classifier leveraging CLIP can achieve strong generalization.

While these works aim at strong detection accuracies, the CLIP latent space itself is somewhat understudied. A better understanding of its topology may further benefit the distinction of synthetic and real images. For example, He *et al.* [18] report that DINOv2 representations of synthetic and real images differ in their sensitivity to noise, which leads to a simple threshold-based classifier. In this work, we investigate the latent space topology of CLIP, and how it relates to generalization capability.

III. Methodology

We investigate three key aspects of the latent space. First, the spatial arrangement of samples from different image sources and their relationship to generalization. Second, the impact of post-processing in the feature space. Third, the interplay between post-processing and image source affiliation.

A. Datasets and CLIP Model

In this Section, we always use 400 images from each of the 7 diffusion-based generators DALL-E 2 [19], DALL-E 3 [20], Stable Diffusion 1.5 [21], SDXL [22], Midjourney 5 [23], Midjourney 6 [23], Adobe Firefly [24], and from 3 real datasets, namely COCO [25], RAISE-6k [26], and LAION [27]. The DALL-E 2 images stem from Corvi *et al.* [11], and we generate all other synthetic images using COCO captions as prompts, with added tokens like "photography" to improve realism. All experiments use representations from the pretrained CLIP ViT-L/14 [1].

B. Spatial Arrangement of Samples in Feature Space

The 768-dimensional CLIP feature space can be qualitatively examined through a dimensionality reduction. Fig. 1 shows a two-dimensional projection of the CLIP spaces of real images and synthetic images via UMAP.Here, real images from the high-quality RAISE dataset are concentrated in a small area, while the lower-quality real images from the LAION and COCO datasets are more widely distributed across the space. The synthetic images are also widely distributed, without distinct clusters. The big differences in the distributions of RAISE compared to LAION and COCO suggest that a classifier trained on RAISE images might face difficulties to generalize to LAION and COCO images. This qualitative impression can be empirically confirmed. We train non-linear SVMs on CLIP features from different training datasets, namely 10k samples from Stable Diffusion paired with 10k samples from RAISE, LAION or COCO. Tab. I shows that SVMs trained on RAISE cannot effectively detect LAION or COCO during testing, and vice versa.

We quantitatively analyze the feature space distribution by computing distances between dataset pairs, distinguishing a source and a target. For each image in the source, we find its nearest neighbor in the target using Euclidean distance in CLIP space. The average of these minimal distances defines the (non-symmetric) divergence between datasets. All divergences are listed in Fig. 2. The non-symmetry shows, for example, for the LAION dataset. LAION has a consistently larger divergence to the other datasets than the other datasets to LAION. This permits the conclusion that LAION is broadly scattered throughout the space. COCO and LAION have larger

Train / Test	COCO/SD	LAION/SD	RAISE/SD	avg
LAION/SD	53.60	92.20	37.20	61.00
RAISE/SD	2.00	0.80	99.60	34.13
COCO/SD	90.40	53.80	15.20	53.13

TABLE I: Accuracies for detecting stable diffusion images with different real datasets: swapping real images from LAION, RAISE, or COCO between training and test generalizes poorly, particularly when training on RISE.

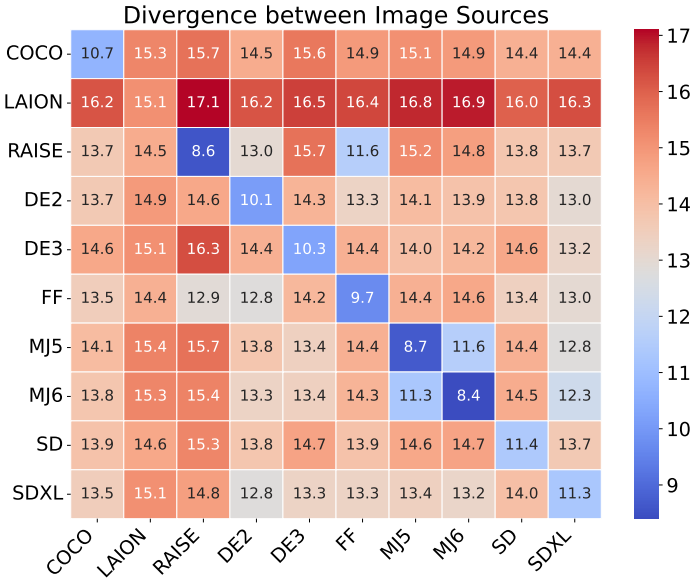
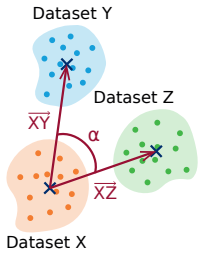


Fig. 2: Divergences (avgerage minimum Euclidean distances between features from one image source to another) of the datasets (see text for details).

$\measuredangle(C, L, X)$	
COCO	-
LAION	-
RAISE	93.9°
DE2	76.9°
DE3	65.2°
FF	84.0°
MJ5	74.9°
MJ6	80.9°
SD	68.6°
SDXL	81.9°



$\measuredangle(X, \cdot, \cdot)$	
COCO	52.7°
LAION	44.1°
RAISE	35.8°
DE2	63.5°
DE3	42.9°
FF	48.5°
MJ5	56.2°
MJ6	55.9°
SD	64.0°
SDXL	69.8°

Fig. 3: Left: angle between dataset axes $\overrightarrow{CL}$ and $\overrightarrow{CX}$ (C: COCO, L: LAION). Middle: sketch of the calculated angles. Right: mean of all angles for all axis pairs that originate from dataset X. Low averages suggest outlier positions of X.

divergences to RAISE than to all synthetic sources, which may also explain the weak generalization between COCO, LAION, and RAISE reported earlier in Tab. I.

The relative location of data sources also has an impact on the difficulty of classification, and the ability to generalize across data sources. We hypothesize that a large angle between the axis from one real to one synthetic image dataset to the axis between one real to another real dataset aids generalization. To investigate this, we computed the centroids for the CLIP representations of each data source. With the centroids of three data sources X, Y, Z, we measure the angle between the vectors $\overrightarrow{XY}$ and $\overrightarrow{XZ}$ as shown in Fig.3.

We test two properties: First, we define the axis $\overrightarrow{CL}$ between COCO and LAION as the baseline for real images. The angle between $\overrightarrow{CL}$ to the axes from COCO to each of the other data sources is reported in Fig. 3 (left). Second, we calculate the average angle of all pairs of axes that originate from a data source X, which is reported in Fig. 3 (right).

Fig. 3 (left) shows that the centroids of most synthetic image sources are positioned almost orthogonally to the COCO-LAION axis. However, also the angle between $\overrightarrow{CL}$ and the axis from COCO to RAISE is almost orthogonal. In combination with the divergences in Fig. 2, this underlines the isolated position of RAISE compared to COCO and LAION.

The right column of Tab. 3 shows substantial variation in average angles, suggesting that some sources are positioned more centrally in CLIP space, while others lie farther away. RAISE and SDXL/SD are notable examples: SDXL/SD's large mean angle of 70° suggests it lies between other sources, whereas RAISE's smaller angle of 36° indicates a more peripheral position. This makes RAISE a less ideal training source, as classifiers may struggle to generalize to other regions of the CLIP space.

C. Impact of Post-Processing

Post-processing (e.g., recompression) oftentimes reduces detector robustness. Hence, to better understand CLIP representations, we also study distances and directions of shifts

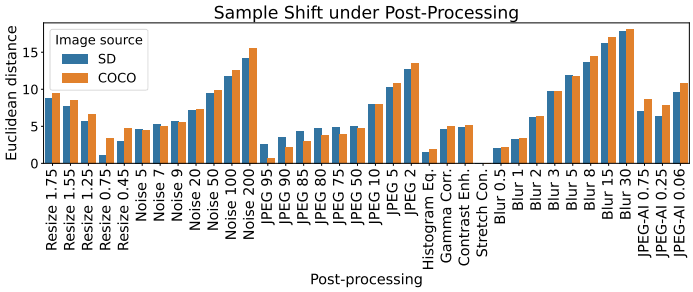


Fig. 4: Euclidean distance of sample shift in latent space after processing of Stable Diffusion and COCO images.

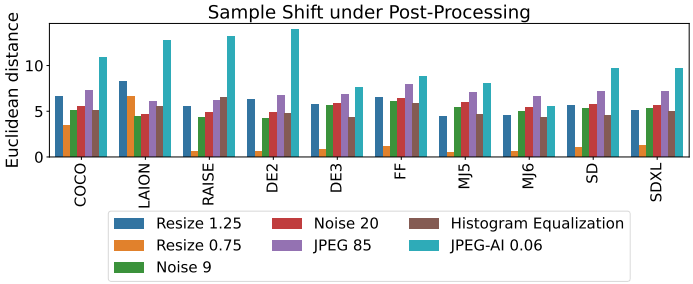


Fig. 5: Comparison of average sample shift measured by Euclidean distance between original and post-processed sample per image source. It can be seen that for example, resizing with factor 0.75 causes a much larger shift for samples from COCO and LAION.

after post-processing. We apply either resizing with factors of $r \in \{0.45, 0.75, 1.25, 1.55, 1.75\}$, Gaussian blur with $\sigma_b = \in \{0.5, 1, 2, 3, 5, 8, 15, 30\}$, JPEG-compression with quality $q_f \in \{95, 90, 85, 80, 75, 50, 10, 5, 2\}$, or additive Gaussian noise with with $\sigma_n \in \{5, 7, 9, 20, 50, 100, 200\}$. Several works locate traces of synthetic images in color properties [13], [28], [29], which is why we also add contrast enhancement, Gamma correction, histogram equalization and contrast stretching, and we also add JPEG AI compression with bitrates of 0.06, 0.25, and 0.75, since JPEG AI introduces similar artifacts as image generators [30] and can disturb detectors [31].

We calculate the Euclidean distances between CLIP embeddings of original and post-processed images. Fig. 4 shows the mean distances per post-processing for COCO and Stable Diffusion images. As expected, higher post-processing strength shifts features by a larger distance. Fig. 5 shows the mean Euclidean distances of 7 example post-processings across all image sources. We observe that the distances may notably vary across image sources. For example, histogram equalization causes larger mean distances on real than on synthetic images. One reason for this observation might be that image generators create higher contrast images due to their optimization for visual aesthetics. Also, resizing affects real images more strongly, while noise affects synthetic images more strongly, which is in line with observations by He *et al.* on the DINOv2 representation [18]. Another interesting property is that the variance of samples within a dataset

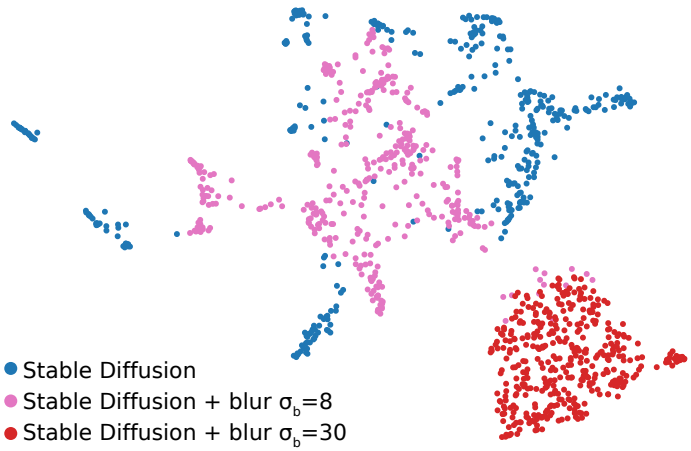


Fig. 6: Distribution of Stable Diffusion samples (blue) originally and after processing with different blur-factors(pink=medium, red=strong). Especially for blur and noise, higher processing factors lead to more dense feature representations.

decreases for high amounts of blur or additive noise, as shown in Fig. 7. This is plausible from a signal processing perspective, when considering that both operations cancel in the infinite limit the distinguishing information in images. A 2-D UMAP projection visualizes this property in Fig. 6 for images from Stable Diffusion, where the distribution contracts for higher amounts of noise.

D. Interplay between Post-Processing and Source Attribution

An interesting question is whether post-processing directly pushes samples across the decision boundary into another class, or whether associated misclassifications stem from another effect. Fig. 5 indicates that a shift due to post-processing may well reach an Euclidean distances of 8. This appears fine when noting that the divergences in Tab. 2 are in the order of about 12 to 20. However, we also calculated the minimal distances between two image sources (i.e., we searched for the two closest samples). These distances range from 5.26 to 13.09 with a mean of 9.15, which indicates that there is the possibility to directly cross a class boundary with post-processing. To further disambiguate these findings, we study the direction of the post-processing shift. Additive noise shifts synthetic samples roughly into the direction of samples from COCO. For example, Midjourney 5 and Midjourney 6 samples with added noise $\sigma = 50$ shift within a cone of approximately a $60°$ angle toward the direction of COCO. JPEG-AI shifts real samples from LAION with a mean angle of about $70°$ towards synthetic samples, where stronger compression rates narrow that angle. Such shifts in the rough direction across the real-synthetic boundary make it more likely that post-processing directly crosses the class boundary.

IV. Application to Classifier Design

CLIP-based classification of synthetic images works well in principle, but it nevertheless leaves room for improvement in

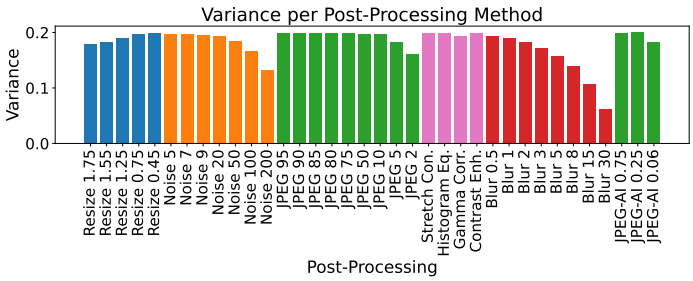


Fig. 7: Feature variance of Stable Diffusion samples after processing.

the generalization, as shown for example in Tab. I in Sec. III-B. We hence loosely adapt and extend the idea by He *et al.* [18] to our findings, and complement the classifier with cues from the post-processing shift of an image.

More specifically, 13 features are calculated from an input image and its post-processed versions. The features consists of 8 Euclidean distances between input and post-processed image, namely from JPEG compression with quality $q \in \{0.85, 0.95\}$, resizing with factors $r \in \{0.75, 1.25\}$, additive noise with $\sigma \in \{9, 20\}$, histogram equalization, and contrast enhancement. The remaining 5 features are cosine similarities between the direction of the mean post-processing shift of the synthetic training images from Stable Diffusion and the input sample. Here, we use noise with $\sigma \in \{9, 20\}$, histogram equalization, gamma correction, and contrast stretching. Classification is done with a SVM with RBF kernel. Different from the previous Section, training is done on 10k images from both the COCO and the Stable Diffusion datasets. Testing is done on 500 "fresh" images per image source that have never been used anywhere else in the paper.

Results are reported in Tab. II for the standard CLIP features as "COCO/SD", only on the post-processing features as "Postproc", and joint training on both features as "COCO/SD + Postproc". The last row fuses the three outputs via majority vote. Overall, post-processing metrics improves the generalization across real datasets, but they perform poorly on synthetic data like DALL-E 3 that behaves similarly to real data. Training on both representations only slightly improves the mean accuracy from 78.28% to 78.42%. However, not all of these errors are correlated, and a majority vote among the three classifiers leads to an average accuracy of 80.52%.

For contextualization we also report performances for methods by Khan *et al.* [5], Ojha *et al.* [2], and Cozzolino *et al.* [4]. Cozzolino *et al.* [4] outperforms our classifier, but we note that the focus of this work is to better understand the CLIP space rather than to optimize performance (and hence also did not, e.g., conduct paired training as Cozzolino *et al.* [4].

V. Conclusion

This work explores the geometry of features in CLIP space and how post-processing influences the distribution of real and synthetic samples. It turns out that different datasets of real and synthetic images occupy different locations in the space,

	COCO	LAION	RAISE	DE2	DE3	FF	MJ5	MJ6	SD	SDXL	avg
Khan *et al.* [5]	69.6	73.0	33.6	6.4	95.0	17.8	65.2	37.6	51.4	65.2	51.48
Cozzolino *et al.* [4]	100.0	95.9	97.9	84.1	1.7	98.4	100.0	98.3	100.0	100.0	87.63
Ojha *et al.* [2]	93.0	92.5	7.0	60.3	65.5	6.1	39.8	9.0	29.0	39.8	44.20
COCO/SD	90.4	53.8	15.2	88.6	64.0	99.8	92.8	81.8	98.6	97.8	78.28
Postproc	98.4	94.4	75.8	45.8	1.4	98.8	82.6	46.0	99.8	100.0	74.30
COCO/SD + Postproc	70.2	98.6	54.6	84.8	15.8	98.2	85.4	78	98.8	99.8	78.42
Fused COCO/SD + Postproc	91.4	96.6	50.4	83.0	15.8	98.8	92.2	77.6	99.6	99.8	80.52

TABLE II: Test accuracies for the CLIP-space metrics. The features show complementary benefits (see text for details).

and we extensively study their relative positions. An improved understanding of these relative locations can aid the construction of more robust classifiers. We demonstrate this with the example of a compact 13-element feature vector of shifts in distances and directions when post-processing an image, which helps to improve the robustness towards the (in CLIP-space) isolated RAISE dataset, while also maintaining robustness to unseen image generators. We hope that this work inspires follow-up works that aim at improved forensic detector design from a characterization of trained representations.

REFERENCES

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2021.

[2] U. Ojha, Y. Li, and Y. Lee, "Towards Universal Fake Image Detectors that Generalize Across Generative Models," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[3] L. Lin, I. Amerini, X. Wang, S. Hu *et al.*, "Robust CLIP-based detector for exposing diffusion model-generated images," in *2024 IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2024, pp. 1–7.

[4] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the Bar of AI-generated Image Detection with CLIP," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4356–4366.

[5] S. A. Khan and D.-T. Dang-Nguyen, "CLIPping the Deception: Adapting Vision-Language Models for Universal Deepfake Detection," in *ACM International Conference on Multimedia Retrieval*, 2024, pp. 1006–1015.

[6] I. Amerini, M. Barni, S. Battiato, P. Bestagini, G. Boato, V. Bruni, R. Caldelli, F. De Natale, R. De Nicola, L. Guarnera *et al.*, "Deepfake Media Forensics: Status and Future Challenges," *MDPI Journal of Imaging*, vol. 11, no. 3, p. 73, 2025.

[7] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros, "CNN-Generated Images Are Surprisingly Easy to Spot... for Now," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8695–8704.

[8] D. Gragnaniello, D. Cozzolino, F. Marra, G. Poggi, and L. Verdoliva, "Are GAN Generated Images Easy to Detect? A Critical Analysis of the State-of-the-Art," in *IEEE International Conference on Multimedia and Expo*, 2021, pp. 1–6.

[9] B. Liu, F. Yang, X. Bi, B. Xiao, W. Li, and X. Gao, "Detecting Generated Images by Real Images," in *European Conference on Computer Vision*, 2022, pp. 95–110.

[10] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei, "Learning on Gradients: Generalized Artifacts Representation for GAN-Generated Images Detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 105–12 114.

[11] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the Detection of Synthetic Images Generated by Diffusion Models," in *International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[12] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "DIRE for Diffusion-Generated Image Detection," arXiv, Tech. Rep. 2303.09295, 2023.

[13] P. He, H. Li, and H. Wang, "Detection of Fake Images via the Ensemble of Deep Representations from Multi Color Spaces," in *IEEE International Conference on Image Processing*, 2019, pp. 2299–2303.

[14] M. Barni, K. Kallas, E. Nowroozi, and B. Tondi, "CNN Detection of GAN-Generated Face Images Based on Cross-Band Co-Occurrences Analysis," in *IEEE International Workshop on Information Forensics and Security)*, 2020, pp. 1–6.

[15] A. Moskowitz, T. Gaona, and J. Peterson, "Detecting AI-Generated Images via CLIP," arXiv, Tech. Rep. 2404.08788, 2024.

[16] D. Cioni, C. Tzelepis, L. Seidenari, and I. Patras, "Are CLIP features all you need for Universal Synthetic Image Origin Attribution?" arXiv, Tech. Rep. 2408.09153, 2024.

[17] S. Smeu, E. Oneata, and D. Oneata, "DeCLIP: Decoding CLIP Representations for Deepfake Localization," arXiv, Tech. Rep. 2409.08849, 2024.

[18] Z. He, P.-Y. Chen, and T.-Y. Ho, "RIGID: A Training-Free and Model-Agnostic Framework for Robust AI-Generated Image Detection," arXiv, Tech. Rep. 2405.20112, 2024.

[19] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical Text-Conditional Image Generation with CLIP Latents," arXiv, Tech. Rep. 2204.06125, 2022.

[20] OpenAI, *DALL-E 3*, 2024, https://openai.com/dall-e-3.

[21] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[22] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis," in *International Conference on Learning Representations*, 2023.

[23] M. Inc., *Midjourney*, 2024, https://www.midjourney.com/home.

[24] Adobe, *Firefly*, 2024, https://www.adobe.com/sensei/generative-ai/firefly.html.

[25] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, 2014, pp. 740–755.

[26] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "RAISE: A Raw Images Dataset for Digital Image Forensics," in *ACM Multimedia Systems Conference*, 2015, pp. 219–224.

[27] C. Schuhmann, R. Vencu, R. Beaumont, R. Kaczmarczyk, C. Mullis, A. Katta, T. Coombes, J. Jitsev, and A. Komatsuzaki, "Laion-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs," arXiv, Tech. Rep. 2111.02114, 2021.

[28] K. Zeng, X. Yu, B. Liu, Y. Guan, and Y. Hu, "Detecting Deepfakes in Alternative Color Spaces to Withstand Unseen Corruptions," in *International Workshop on Biometrics and Forensics*, 2023, pp. 1–6.

[29] H. Li, B. Li, S. Tan, and J. Huang, "Identification of Deep Network Generated Images Using Disparities in Color Components," *Signal Processing*, vol. 174, p. 107616, 2020.

[30] S. Bergmann, D. Moussa, F. Brand, A. Kaup, and C. Riess, "Forensic Analysis of AI-Compression Traces in Spatial and Frequency Domain," *Pattern Recognition Letters*, vol. 180, pp. 41–47, 2024.

[31] E. D. Cannas, S. Mandelli, N. Popovic, A. Alkhateeb, A. Gnutti, P. Bestagini, and S. Tubaro, "Is JPEG AI going to change image forensics?" arXiv, Tech. Rep. 2412.03261, 2024.