AI-Generated Video Detection via Spatio-Temporal Anomaly Learning, PRCV 2024

Preprint · October 2024

DOI: 10.1007/978-981-97-8792-0_32

4 authors:

Jianfa Bai
Communication University of China
1 PUBLICATION **3** CITATIONS
SEE PROFILE

Man Lin
Communication University of China
5 PUBLICATIONS **16** CITATIONS
SEE PROFILE

Gang Cao
Communication University of China
68 PUBLICATIONS **1,525** CITATIONS
SEE PROFILE

Zijie Lou
Communication University of China
8 PUBLICATIONS **23** CITATIONS
SEE PROFILE

AI-Generated Video Detection via Spatial-Temporal Anomaly Learning

Jianfa Bai[1,2], Man Lin[1,2], Gang Cao[1,2(✉)], and Zijie Lou[1,2]

[1] School of Computer and Cyber Sciences, Communication University of China, Beijing 100024, China
[2] State Key Laboratory of Media Convergence and Communication, Communication University of China, Beijing 100024, China
{jianfa, lyan924, gangcao, louzijie2022}@cuc.edu.cn

Abstract. The artificial intelligence (AI)-generated videos become more and more realistic with the advancement of generation models. Such synthetic videos are indistinguishable from the real ones by human eyes, and may be easily created by malicious users to spread false information. To prevent the misuse, we propose an effective AI-Generated Video Detection (AIGVDet) scheme with spatial-temporal convolutional neural network (CNN) and decision fusion strategy. Specifically, two separate ResNet detectors are learned for identifying the anomalies in spatial and optical flow domains, respectively. To enhance the discrimination ability of AIGVDet, the frame-level prediction results of such two detectors are aggregated to the final video detection result based on multiple stages decision fusion. A new large-scale generated video dataset (GVD) is created as a benchmark for network training and evaluation. Extensive experimental results verify the high generalization ability and robustness of our AIGVDet scheme in detecting AI-generated videos blindly. The code and dataset are available at https://github.com/multimediaFor/AIGVDet.

Keywords: Video forensics · Generated video detection · Spatial-temporal anomaly · Optical flow · Decision fusion

1 Introduction

Recent development of large models has greatly promoted the advancement of AI-generated content [25]. The AI-generated videos with exceptional quality, rapid creation and cost-effectiveness are revolutionizing industries, such as short and long-form video production, gaming and advertising. However, there also appear high risks associated with such videos including the spread of misinformation and manipulation of public opinion. Many generated videos are so realistic that they are virtually indistinguishable from real ones, particularly with the latest generation models like Sora [19]. Despite regulatory attempts such as Biden's signing of AI act [21], reliable blind detection tools are still necessary to differentiate between the AI-generated and real videos.

J. Bai and M. Lin are Contributed equally.

Z. Lin et al. (Eds.): PRCV 2024, LNCS 15040, pp. 460–470, 2025.
https://doi.org/10.1007/978-981-97-8792-0_32

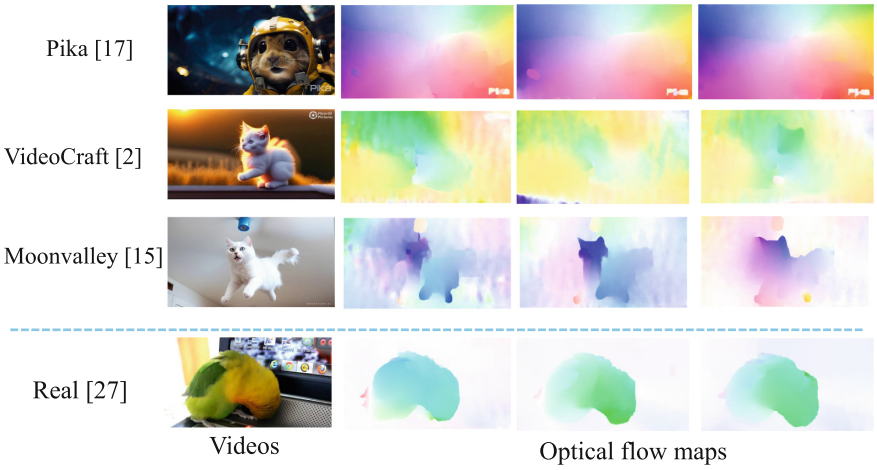

Fig. 1. The first frame and the first three optical flow maps of the videos generated by three AI models, along with those of a real video. The hue and saturation in optical flow maps indicate the direction and magnitude of pixel displacements between two neighboring frames.

Currently, available algorithms for distinguishing between generated and real videos primarily consist of methods used for detecting generated images [3,4,8,14,22–24] and those used for detecting forged face videos [1,9,27]. There are no dedicated methods specifically designed for detecting generated videos. To address such a gap in existing works, we formally point out the AI-generated video detection problem and propose an effective solution. We observe that low-quality generated videos may exhibit visual anomalies, such as abnormal textures and violation of physical rules. High-quality generated videos, which are indistinguishable from real ones to the naked eye, are likely to manifest temporal discontinuities. Figure 1 illustrates the optical flow maps computed from example synthetic and real videos. Despite the remarkable visual fidelity of generated video frames, their optical flow maps exhibit less smoothness and blurry contours compared to those of the real videos. To capture such disparities, a simple yet effective AI-generated video detection (AIGVDet) model is proposed. The RGB frames and their corresponding optical flow maps serve as inputs, with a two-branch ResNet50 [10] encoder thoroughly exploring abnormalities in both modalities. In the end, a decision-level fusion binary classifier is assembled to effectively integrate information and enhance the model's discriminative ability. Additionally, a large-scale generated video benchmark dataset is created for network training and evaluation, comprising synthetic videos from 11 different generator models. Extensive experiments showcase the generalization capability and robustness of our proposed detector.

To sum up, the main contributions of this work can be concluded as the following two aspects:

• We propose an effective scheme to distinguish between generated videos and real ones, leveraging inconsistencies in both spatial and temporal domains.
• We introduce a dataset containing 11,618 generated videos and conduct experiments on generalization and robustness, yielding significant results.

2 Related Work

2.1 AI-Generated Image Detection

As a new digital forensics problem, there are no prior works on the blind detection of AI-generated videos. Correlatively, many forensic methods [3,4,8,14,22–24] have been proposed for detecting AI-generated images. In [8,22], large-scale training samples and data augmentation strategies are employed to learn the common artifacts in GAN generated images, resulting in detectors with good generalization. On the other hand, some works [3,23] focus on the detection of images generated by diffusion models. In [4,24], the detection of images generated by both GAN and diffusion models is achieved based on the contrastive language-image pre-training model (CLIP) [18]. However, as found in our experiments, such detectors perform poorly in detecting the generated videos. This can be attributed to the different generation mechanisms between synthetic images and videos, such as spatial-temporal consistency.

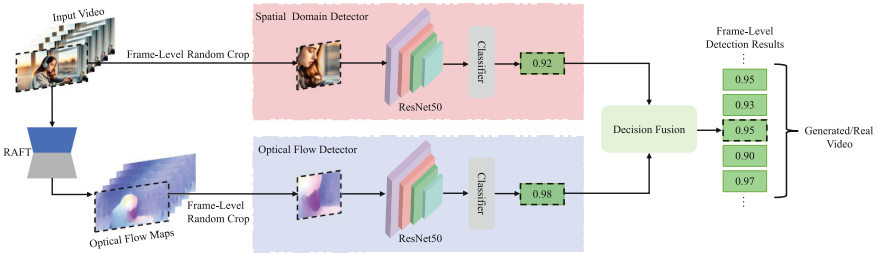

Fig. 2. Overall pipeline of the proposed generated video detection scheme AIGVDet, where RAFT [20] is the method for calculating optical flow maps.

2.2 Forged Face Video Detection

There are also many prior works [1,9,27] in detecting forged face videos. Yang et al. [27] exploit the appearance and motion characteristics of lips to defend against sophisticated deepfake attacks. Gu et al. [9] design different modules to capture the inconsistencies caused by subtle movements within and between densely sampled video snippets. Such methods focus on capturing physiological defects and subtle facial abnormalities, which are not included in the widespread generated videos without human faces.

3 Proposed AIGVDet Scheme

3.1 Two-Branch Spatial-Temporal Detector: AIGVDet

The overall pipeline of our AIGVDet scheme is illustrated in Fig. 2. It comprises two individually trained detectors both with ResNet50 [10] backbone network, namely the spatial domain and the optical flow detectors. The former explores the abnormality of spatial pixel distributions (e.g. texture, noise, etc.) within single RGB frames, while the latter captures temporal inconsistencies via optical flow. Let a video be denoted by N frames $\{I_i\}_{i=1}^{N} \in \mathbb{R}^{W\times H\times 3}$ with a spatial resolution of $W \times H$ pixels. Then the optical flow maps $F_i \in \mathbb{R}^{W\times H\times 3}$ between adjacent frames are calculated by the RAFT [20] estimator $\mathcal{F}(\cdot)$ as

$$F_i = \mathcal{F}(I_i,\ I_{i+1}). \tag{1}$$

Each frame I_i and its corresponding optical flow F_i are then randomly cropped to 448×448, and fed into the spatial domain encoder $\mathrm{R}(\cdot)$ and the optical flow branch encoder $\mathrm{O}(\cdot)$, respectively. Subsequently, the extracted two-modal features V_I^i and V_F^i are sent to two binary classifiers, both consist of a global average pooling layer $\mathrm{GAP}(\cdot)$, a fully connected layer $\mathrm{FC}(\cdot)$ and a sigmoid activation function $\mathrm{Sigmoid}(\cdot)$. This process maps the two feature vectors to probabilities P_I^i and P_F^i, indicating the likelihood of being generated. That is,

$$V_I^i = \mathrm{R}(I_i), \quad V_F^i = \mathrm{O}(F_i). \tag{2}$$

$$P_I^i = \mathrm{Sigmoid}(\mathrm{FC}(\mathrm{GAP}(V_I^i))),$$

$$P_F^i = \mathrm{Sigmoid}(\mathrm{FC}(\mathrm{GAP}(V_F^i))). \tag{3}$$

To integrate the representational capabilities of these two feature vectors, decision-level fusion is employed to obtain the prediction result P^i for each frame. It is the weighted summation of P_I^i and P_F^i as

$$P^i = \alpha P_I^i + (1-\alpha)\, P_F^i, \tag{4}$$

where $\alpha \in (0,1)$ is a weight to balance the predictions from spatial domain and optical flow modalities. Lastly, the final video-level prediction result P is computed as

$$P = \frac{1}{N-1}\sum\nolimits_{i=1}^{N-1} P^i. \tag{5}$$

Here, $P \in (0,1)$ signifies the probability of being an AI-generated video.

Note that the state-of-the-art video generation algorithms still encounter some technical bottlenecks in realistic generation. Due to a lack of deep understanding of physical laws, the generated content may deviate from real-world physics. There may exist discrepancies in the direction and speed of object motion compared to reality. Such findings motivate us to leverage optical flow maps to enhance the discrimination of generated videos. The optical flow is an

estimation of the per-pixel movement between neighboring frames. We employ the RAFT [20] algorithm to compute the optical flow prediction between adjacent frames. It comprises three main components, i.e., feature extraction, computing visual similarity, and iterative updates [20].

3.2 Generated Video Dataset

To train and evaluate the AIGVDet model, the GVD is constructed by collecting 11,618 video samples yielded by 11 state-of-the-art generator models. Each generator model is trained on a distinct real video dataset tailored to its specific task. The most two common types of generation models, i.e., Text-to-Video (T2V) and Image-to-Video (I2V) are involved specifically. T2V refers to the automatic generation of corresponding videos based on the content described in the text. I2V refers to generating videos using images accompanied by descriptive information, or just images [26]. Specific details of the GVD are presented in Table 1. The main collection source 'Discord' [6] is a free network communication and digital distribution platform. Within the video-generating clubs of such a platform, users can share and showcase their videos synthesized using various generator models.

Table 1. Details of our collected generated video dataset (GVD).

Type	Name	Number	Resolution	Format	Source
T2V	Moonvalley [15]	3.55k	1184×672	MP4	Discord
	VideoCraft [2]	1.5k	1024×576	MP4	Discord
	Pika [17]	1.0k	1024×576-1088×640	MP4	Discord
	NeverEnds [16]	1.0k	1024×576	MP4	Discord
	Emu [7]	900	512×512	MP4	Discord
	VideoPoet [13]	120	512×896	MP4	Official Web
	Hotshot [11]	500	672×384	GIF	Official Web
	Sora [19]	48	512×512-1920×1088	MP4	Official Web
I2V	Moonvalley [15]	1.0k	626×626-784×1184	MP4	Discord
	Pika [17]	1.0k	640×832-1152×640	MP4	Discord
	NeverEnds [16]	1.0k	512×960-1024×576	MP4	Discord

4 Experiments

4.1 Experimental Setup

The spatial domain detector and the optical flow detector are trained separately. All training steps are conducted in the same way apart from the difference in input. Further details regarding datasets, preprocessing methods, and evaluation metrics are elaborated below.

Datasets. To comprehensively explore the generalization of our AIGVDet, training is conducted solely on generated videos from a single generation model, while testing on videos from various sources to simulate real-world scenarios. Specifically, we utilize 550 T2V generated videos from the Moonvalley [15] and 550 real videos from the YouTube_vos2 dataset [28] for model training and validation, with a training-validation set ratio of 10:1. Each video is sampled 95 RGB frames and their corresponding 94 optical flow maps. All generated videos in GVD, except for those used for model training and validation, will be utilized for testing. The corresponding number of real test videos is sourced from the GOT dataset[12].

Preprocessing. Since some collected generated videos contain watermarks at the bottom, the frames are firstly cropped to remove such watermarks to avoid bias. During training, considering that the frames of real videos in the training dataset are all in JPEG format, we randomly compress the generated frames with a JPEG compression factor ranging from 70 to 90 to mitigate the impact of compression. Before being fed into the detectors, all input images are cropped to a size of 448×448. Random cropping is employed during training, while central cropping is utilized during testing.

Comparative Methods and Metrics. Comparable detectors specifically designed for generated videos are lacking, thus we compare against two pioneering and classic detectors for generated images, namely Wang [22] and DIRE [23]. We evaluate the Wang [22] and DIRE [23] methods in two ways: without retraining and retraining on our training samples. The training and testing strategies remain consistent with their original settings. Wang [22]utilizes RGB images as input, which are resized to 256x256, and then cropped to 224x224. DIRE [23], on the other hand, adds noise to RGB images, denoises them, and then applies the same processing steps as Wang [22]. Accuracy (ACC) and area under the receiver-operating curve (AUC) are used as the video-level evaluation metrics. The video-level classification threshold is set at 0.5 and the weight α during decision-level fusion is set at 0.5.

Table 2. Ablation test results. ACC (%) and AUC (%) among different variants of proposed scheme on T2V generated videos.

Variants	Moonvalley	VideoCraft	Pika	NeverEnds	Average
$S_{spatial}$	98.7/**100**	80.4/93.9	**80.3**/93.7	76.9/92.0	84.1/94.9
$S_{optical}$	98.2/99.8	81.1/95.0	68.5/92.7	77.1/92.6	81.2/95.0
$S_{optical\_no\_cp}$	98.2/99.8	79.7/94.4	62.5/89.8	75.6/91.7	79.0/93.9
FF_{concat}	**99.7**/**100**	67.2/92.2	64.5/92.2	69.8/94.1	75.3/94.6
FF_{add}	97.3/**100**	52.3/90.4	50.4/87.6	52.4/89.4	63.1/91.9
AIGVDet	99.5/**100**	**83.6**/**96.8**	79.9/**95.7**	**79.1**/**95.4**	**85.5**/**97.0**

Implementation. The proposed method is implemented using the PyTorch deep learning framework, and all experiments are conducted on an A800 GPU. Binary cross-entropy is used as the loss function. The backbone ResNet50 [10] is pre-trained with ImageNet [5]. We use the Adam optimizer as the optimization function with an initial learning rate 1e-4. The learning rate is reduced by a factor of 10 if the validation accuracy does not increase after 5 epochs. Training is terminated when the learning rate reaches 1e-6. Data augmentation is applied to 10% of training samples, includes Gaussian blurring with a sigma of 0.5, JPEG compression with a compression factor of 75, and random flipping.

4.2 Ablation Study

Multiple ablation experiments were conducted to analyze the factors contributing to the cross-model generalization of this scheme. For this purpose, we trained multiple variants using the same settings as stated above and assessed their performance on the subsets of T2V generated videos. These variants include: trained solely on RGB frames ($S_{spatial}$), trained solely on optical flow maps ($S_{optical}$), trained solely on optical flow maps without cropping operation ($S_{optical\_no\_cp}$), trained simultaneously on RGB frames and optical flow maps with feature fusion through concatenation (FF_{concat}), trained simultaneously on RGB frames and optical flow maps with feature fusion through element-wise addition (FF_{add}). The results of comparative experiments are shown in Table 2.

Efficacy of Feature Representation. Comparing the results of $S_{spatial}$ and $S_{optical}$ with our proposed AIGVDet, it can be observed that combining

Table 3. ACC (%) and AUC (%) comparison of different detectors on different test datasets. †: retrained version with our training dataset.

Datasets		Detectors				
Type	**Name**	**Wang[22]**	**Wang†[22]**	**DIRE[23]**	**DIRE†[23]**	**AIGVDet**
T2V	Moonvalley [15]	50.0/39.0	**99.7**/**100**	50.7/66.7	98.9/99.9	99.5/**100.0**
	VideoCraft [2]	50.1/42.9	69.5/92.3	49.6/48.4	72.4/90.6	**83.6**/**96.8**
	Pika [17]	50.2/42.0	64.1/89.5	53.9/60.0	71.3/89.2	**79.9**/**95.7**
	NeverEnds [16]	50.1/28.3	76.0/95.5	49.8/49.4	81.6/93.5	**79.1**/**95.4**
	Emu [7]	50.0/29.1	94.4/**99.7**	49.3/41.5	**96.7**/99.6	93.7/99.1
	VideoPoet [13]	50.0/64.0	68.3/95.6	50.0/34.1	69.6/84.8	**76.3**/**97.6**
	Hotshot [11]	50.0/13.9	52.8/81.5	53.7/55.3	59.5/75.0	**65.7**/**97.9**
	Sora [19]	50.0/70.4	55.2/89.0	50.0/56.9	56.3/81.8	**68.8**/**93.2**
	Average	50.1/41.2	72.5/92.9	50.9/51.5	75.8/89.3	**80.8**/**97.0**
I2V	Moonvalley [15]	50.1/28.5	78.8/94.3	50.6/46.8	82.3/94.3	**83.3**/**96.1**
	Pika [17]	50.1/37.0	76.9/93.8	51.2/47.9	82.8/93.6	**88.3**/**96.5**
	NeverEnds [16]	50.0/33.8	68.5/92.5	50.3/46.4	73.1/90.1	**76.7**/**96.0**
	Average	50.1/33.1	74.7/93.5	50.7/47.0	79.4/92.7	**82.8**/**96.2**

features from the spatial domain and the optical domain yields better results than using only a single spatial or optical branch.

Impact of Preprocessing Methods. When directly training and testing the detector with entire optical flow maps without cropping, $S_{optical\_no\_cp}$ shows a 2.2% decrease in average detection accuracy compared to $S_{optical}$. The cropping operation enhances the detector's attention to local subtle alterations in the optical flow maps while simultaneously reducing interference from global information. This suggests that local information is more suitable for detecting generated videos than global information.

Impact of Fusion Methods. The RGB frames carry more visual details than optical flow maps, which also compensate for temporal information not contained in the RGB frames. Therefore, we study how to better fuse such two types of information. The results reveal that the decision fusion adopted by AIGVDet is more effective than the feature-level fusions via concatenation FF_{concat} and direct addition FF_{add}. The detection network schemes based on feature fusion tend to overfit on the Moonvalley sub-dataset of T2V dataset and perform badly on the other ones.

4.3 Assessment of Generalization Ability

The varieties and quantities of video generation models will continue to evolve and improve, making the generalization performance of detectors crucial in real-world scenarios. Table 3 presents the detection results comparing the performance of Wang [22] and DIRE [23] methods, both without retraining and after retraining, with our proposed AIGCDet method.

Firstly, for T2V generation models, the AIGVDet achieves an ACC of 99.5% and an AUC of 100% on the Moonvalley [15] dataset, which is consistent with the training set. It also achieves good results on other unseen models, especially the latest Sora [19], reaching an AUC of 93.2%. Even for generated videos of different resolutions such as Videocraft [2] (1024×576), Emu [7] (512×512), Hotshot [11] (672×384), and VideoPoet [13] (512×896), effective discrimination can also be achieved. On average, the detection accuracy of AIGVDet (80.8%) is significantly higher than that of Wang† (72.5%) and DIRE† (75.8%). Additionally, the ACC and AUC of Wang and DIRE without retraining are extremely low, close to 50%. The poor performance of these generated image detectors can be attributed to the neglect of temporal anomaly information.

Our AIGVDet scheme also performs well on the more challenging I2V generated videos. Compared with T2V, the I2V videos exhibit less pronounced motion with only subtle local changes, such as the ripples on water surface. Moreover, I2V videos are primarily generated using real image inputs, which further incurs challenges for detection. Nonetheless, the average ACC (82.8%) and AUC (96.2%) demonstrate that our detector can generalize well to different types and resolutions of unseen generated videos.

4.4 Robustness Evaluation

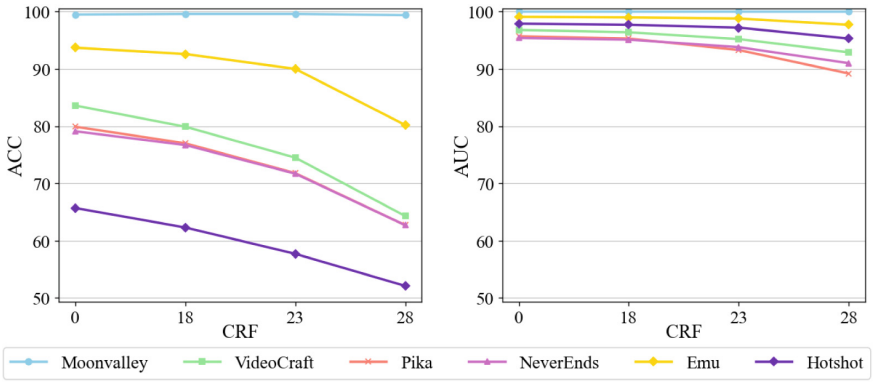

Fig. 3. Robustness evaluation results against post H.264 compression with different quality factors (CRFs).

The robustness against video compression, which is a common post-processing method in real-life scenarios, is evaluated. CRF is a parameter that controls the compression quality of H.264. We test compression factors with CRF = 0, 18, 23, and 28 (0 indicates no compression). Video recompression is applied to both generated and real videos. We conduct this experiment on T2V videos from Moonvalley [15], VideoCraft [2], Pika [17], NeverEnds [16], Hotshot [11] and Emu [7]. The results in Fig. 3 show that the AUC and ACC decreases to some extent with the increase of compression degree. However, the AUC consistently exceeds 88%.

5 Conclusion

In this paper, a simple yet effective scheme is proposed to detect AI-generated videos. Leveraging our newly constructed GVD, the proposed AIGVDet effectively captures and integrates spatial-temporal inconsistencies present in RGB frames and optical flow maps to distinguish between generated and authentic videos. Experimental results indicate that better performance can be achieved by employing separate training of spatial and temporal detectors, followed by decision fusion of the predicted results, compared to other methods. Moreover, our detection scheme exhibits good generalization to videos generated by various unknown generator models and is also effective against video compression. We aim for our work to serve as a robust baseline for detecting generated videos.

Acknowledgments. Supported by National Natural Science Foundation of China (62071434), Fundamental Research Funds for the Central Universities (CUC24GT01), CUC Public Computing Cloud. Jianfa Bai and Man Lin are Contributed equally.

References

1. Caldelli, R., Galteri, L., Amerini, I., Del Bimbo, A.: Optical flow based CNN for detection of unlearnt deepfake manipulations. Elsevier Pattern Recognit. Lett. **146**, 31–37 (2021)
2. Chen, H., Xia, M., He, Y., Zhang, Y., Cun, X., Yang, S., Xing, J., Liu, Y., Chen, Q., Wang, X., et al.: Videocrafter1: Open diffusion models for high-quality video generation (2023). arXiv: 2310.19512
3. Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: On the detection of synthetic images generated by diffusion models. In: IEEE International Conference on Acoustics, Speech and Signal Processing, pp. 1–5 (2023)
4. Cozzolino, D., Poggi, G., Corvi, R., Nießner, M., Verdoliva, L.: Raising the bar of ai-generated image detection with clip (2023). arXiv:2312.00195
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: a large-scale hierarchical image database. In: IEEE Conference on computer Vision and Pattern Recognition, pp. 248–255 (2009)
6. Discord: https://discord.com/ (2023)
7. Girdhar, R., Singh, M., Brown, A., et al.: Emu video: Factorizing text-to-video generation by explicit image conditioning (2023). arXiv: 2311.10709
8. Gragnaniello, D., Cozzolino, D., Marra, F., Poggi, G., Verdoliva, L.: Are gan generated images easy to detect? a critical analysis of the state-of-the-art. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2021)
9. Gu, Z., Chen, Y., Yao, T., Ding, S., Li, J., Ma, L.: Delving into the local: Dynamic inconsistency learning for deepfake video detection. In: AAAI Conference on Artificial Intelligence, vol. 36, pp. 744–752 (2022)
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
11. Hotshot: https://www.hotshot.co/ (2023)
12. Huang, L., Zhao, X., Huang, K.: Got-10k: a large high-diversity benchmark for generic object tracking in the wild. IEEE Trans. Pattern Anal. Mach. Intell. **43**(5), 1562–1577 (2019)
13. Kondratyuk, D., Yu, L., Gu, X., Lezama, J., Huang, J., Hornung, R., Adam, H., Akbari, H., Alon, Y., Birodkar, V., et al.: Videopoet: A large language model for zero-shot video generation (2023). arXiv: 2312.14125
14. Lou, Z., Cao, G., Lin, M.: Black-box attack against GAN-generated image detector with contrastive perturbation. Elsevier Eng. Appl. Artif. Intell. **124**, 106594 (2023)
15. Moonvalley: https://moonvalley.ai/ (2023)
16. NeverEnds: https://neverends.life (2023)
17. Pika: https://www.pika.art/ (2023)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning, pp. 8748–8763. PMLR (2021)
19. openai sora: https://openai.com/sora (2024)
20. Teed, Z., Deng, J.: Raft: recurrent all-pairs field transforms for optical flow. In: Springer European Conference on Computer Vision, pp. 402–419 (2020)
21. The Washington Post: https://www.washingtonpost.com/technology/2023/10/30/biden-artificial-intelligence-executive-order/ (2023)

22. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: CNN-generated images are surprisingly easy to spot... for now. In: IEEE International Conference on Computer Vision and Pattern Recognition, pp. 8695–8704 (2020)
23. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection (2023). arXiv: 2303.09295 . arXiv: 2303.09295
24. Wu, H., Zhou, J., Zhang, S.: Generalizable synthetic image detection via language-guided contrastive learning (2023). arXiv:2305.13800
25. Wu, J., Gan, W., Chen, Z., Wan, S., Lin, H.: Ai-generated content (aigc): a survey (2023). arXiv:2304.06632
26. Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., Jiang, Y.G.: A survey on video diffusion models (2023). arXiv:2310.10647
27. Yang, C.Z., Ma, J., Wang, S., Liew, A.W.C.: Preventing deepfake attacks on speaker authentication by dynamic lip movement analysis. IEEE Trans. Inf. Forensics Secur. **16**, 1841–1854 (2020)
28. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: IEEE International Conference on Computer Vision, pp. 5188–5197 (2019)