Davide Cozzolino, Koki Nagano, Lucas Thomaz, Angshul Majumdar, and Luisa Verdoliva

Synthetic Image Detection

Highlights from the IEEE Video and Image Processing Cup 2022 Student Competition

The Video and Image Processing (VIP) Cup is a student competition that takes place each year at the IEEE International Conference on Image Processing (ICIP). The 2022 IEEE VIP Cup asked undergraduate students to develop a system capable of distinguishing pristine images from generated ones. The interest in this topic stems from the incredible advances in the artificial intelligence (AI)-based generation of visual data, with tools that allow the synthesis of highly realistic images and videos. While this opens up a large number of new opportunities, it also undermines the trustworthiness of media content and fosters the spread of disinformation on the Internet. Recently, there has been strong concern about the generation of extremely realistic images by means of editing software that includes the recent technology on diffusion models [1], [2]. In this context, there is a need to develop robust and automatic tools for synthetic image detection.

In the literature, there has been an intense research effort to develop effective forensic image detectors, and many of them, if properly trained, appear to provide excellent results [3]. Such results, however, usually refer to ideal conditions and rarely stand the challenge of real-world application. First of all, testing a detector on images generated by the very same models seen in the training phase leads to overly optimistic results. In fact, this is not a realistic scenario. With the evolution of technology, new architectures and different ways of generating synthetic data are continuously proposed [4], [5], [6], [7], [8]. Therefore, detectors trained on some specific sources will end up working on target data of a very different nature, often with disappointing results. In these conditions, the ability of generalizing to new data becomes crucial to keep providing a reliable service. Moreover, detectors are often required to work on data that have been seriously impaired in several ways. For example, when images are uploaded on social networks, they are normally resized and compressed to meet internal constraints. These operations tend to destroy important forensic traces, calling for detectors that are robust to such events and degrade performance gracefully. To summarize, to operate successfully in the wild, a detector should be robust to image impairments and, at the same time, able to generalize well on images coming from diverse and new models.

In the scientific community, there is still insufficient (although growing) awareness of the centrality of these aspects in the development of reliable detectors. Therefore, we took the opportunity of this VIP Cup to push further along this direction. In designing the challenge, we decided to consider an up-to-date, realistic setting with test data including 1) both fully synthetic and partially manipulated images and 2) images generated by both established generative adversarial network (GAN) models and newer architectures, such as diffusion-based models. With the first dichotomy, we ask that the detectors be robust to the occurrence of images that are only partially synthetic, thus with limited data on which to base the decision. As for architectures, there is already a significant body of knowledge on the detection of GAN-generated images [9], but new text-based diffusion models are now gaining the spotlight, and generalization becomes the central issue. With the 2022 IEEE VIP Cup, we challenged teams to design solutions that are able to work in the wild as only a fraction of the generators used in the test data are known in advance.

In this article, we present an overview of this challenge, including the competition setup, the teams, and their technical approaches. Note that all of the teams were composed of a professor, at most one graduate student (tutor), and undergraduate students (from a minimum of three to a maximum of 10 students).

Tasks, resources, and evaluation criteria

Tasks

The challenge consisted of two phases: an open competition (split into two parts), in which any eligible team could participate, and an invitation-only final. Phase 1 of the open competition was designed to provide teams with a simplified version of the problem at hand to familiarize themselves with the task, while phase 2 was designed to tackle a

Digital Object Identifier 10.1109/MSP.2023.3294720
Date of current version: 3 November 2023

more challenging task: synthetic data generated using architectures not present in the training. The synthetic images included in phase 1 were generated using five known techniques, while the generated models used in phase 2 were unknown. During the final competition, the three highest-scoring teams from the open competition were selected and were allowed to provide another submission graded on a new test set. Information about the challenge is also available at https://grip-unina.github.io/vipcup2022/.

Resources

Participants were provided with a labeled training dataset of real and synthetic images. In particular, the dataset available for phase 1 comprised real images from four datasets (FFHQ [4], Imagenet [17], COCO [18], and LSUN [19]), while synthetic images were generated using five known techniques: StyleGAN2 [11], StyleGAN3 [12], GLIDE [5], Taming Transformers [10], and inpainted images with Gated Convolution [13]. All the images of the test data were randomly cropped and resized to 200 × 200 pixels and then compressed using JPEG at different quality levels. This pipeline was used to simulate a realistic scenario where images were randomly resized and compressed as happens when they are uploaded to a social network. In addition, they all had the same dimensions to avoid leaking information on the used generators (some models only generate data at certain specific resolutions). Some examples of generated images used during the competition are shown in Figure 1.

Teams were provided with Python scripts to apply these same operations to the training dataset. For phase 2, there were no available datasets since the generated models in this case were unknown to the teams. However, participants were free to use any external data, besides the competition data. In addition, participants were allowed to use any available state-of-the-art methods and algorithms to solve the problems of the challenge.

Teams were requested to provide the executable code to the organizers to test the algorithms on the evaluation datasets. The Python code was executed inside a Docker container with a GPU of 16 GB with a time limit of one hour to process a total of 5,000 images. The teams were allowed to submit their code and evaluate their performance five times during the period from 8 August to 5 September 2022.

Evaluation criteria

The submitted algorithms were scored by means of balanced accuracy for the detection task (score = 0.7 × accuracy phase 1 + 0.3 × accuracy phase 2). The three highest-scoring teams from the open competition stage were selected as finalists. These teams had the opportunity to make an additional submission on 8 October on a new dataset and were invited to compete in the final stage of the challenge at ICIP 2022 on 16 October 2022 in Bordeaux. Due to some travel limitations, on that occasion, they could make a live or prerecorded presentation, followed by a round of questions from a technical committee. The event was hybrid to ensure a wide participation and allow teams who had visa issues to attend virtually. In the final phase of the challenge, the judging committee considered the following parameters for the final evaluation (maximum score was 12 points):

- the innovation of the technical solution (one to three points)
- the performance achieved in phase 1 of the competition, where only known models were used to generate synthetic data (one to three points)
- the performance achieved in phase 2 of the competition, where unknown

FIGURE 1. Examples of synthetic images from the datasets used in the open competition. The first row shows samples from GLIDE [5], Taming Transformers [10], StyleGAN2 [11], StyleGAN3 [12], and inpainting with Gated Convolution [13]. The second row shows samples from BigGAN [14], DALL-e mini [6], Ablated Diffusion Model [15], Latent Diffusion [7], and LaMa [16]. The images in the fifth column are only locally manipulated (the regions outlined in red are synthetic).

models were used to generate synthetic data (one to three points)
- the quality and clarity of the final report, a four-page full conference paper in the IEEE format (one to three points)
- the quality and clarity of the final presentation (either prerecorded or live), a 15-min talk (one to three points).

2022 VIP Cup statistics and results

The VIP Cup was run as an online class through the Piazza platform, which allowed easy interaction with the teams. In total, we received 82 registrations for the challenge, 26 teams accessed the Secure CMS platform, and 13 teams made at least one valid submission. Teams were from 10 different countries across the world: Bangladesh, China, Germany, Greece, India, Italy, Poland, Sri Lanka, United States of America, and Vietnam.

Figure 2 presents the accuracy results obtained by the 13 teams participating in the two phases of the open competition. First, we can observe that the performance on test set 1 including images from known generators was much higher than those obtained in an open set scenario, where generators are unknown. More specifically, there were accuracy drops of around 10% for the best techniques, confirming the difficulty to detect synthetic images coming from unknown models. Then, we noted that even for the simpler scenario, only four teams were able to achieve an accuracy above 70%, which highlights that designing a detector that can operate well on both fully and locally manipulated images is not an easy task.

In Figure 3, we present some additional analyses of all of the submitted algorithms. Figure 3(a) aims at understanding how much computational complexity (measured by the execution time to process 10,000 images) impacts the final score. Interestingly, there is only a weak correlation between computation effort and performance, with methods that achieve the same very high score (around 90%) with very different execution times. Figure 3(b), instead, shows the results of each method on test set 1 and test set 2. In this case, a strong correlation is observed: if an algorithm performs well/poorly on test set 1, the same happens on test set 2, even if the datasets do not overlap and are completely separated in terms of generating models.

Finally, in Figure 4, we study in some more detail the performance of the three best performing techniques, reporting the balanced accuracy for each method on each dataset. For test set 1 (known models), the most difficult cases are those involving local manipulations. The same holds for test set 2 (unknown models) with the additional problem of images fully generated using diffusion models, where performances are on average lower than those obtained on images created by GANs. We also provide results in terms of area under the receiver operating characteristic curve in Figure 5. In this situation, we can note that the first and second places reverse on test set 2, which underlines the importance to properly set the right threshold for the final decision. A proper choice of the validation set is indeed very important to carry out a good calibration.

Highlights of the technical approaches

In this section, we present an overview of the approaches proposed by all of the participating teams for the challenge. All proposed methods relied on learning-based approaches and train deep neural networks on a large dataset of real and synthetic images. Many diverse architectures were considered: GoogLeNet, ResNet, Inception, Xception, DenseNet, EfficientNet, MobileNet, ResNeXt, ConvNeXt, and the more recent Vision Transformers. The problem was often treated as a binary classification task (real versus fake), but some teams approached it as a multiclass classification problem with the aim to increase the degrees of

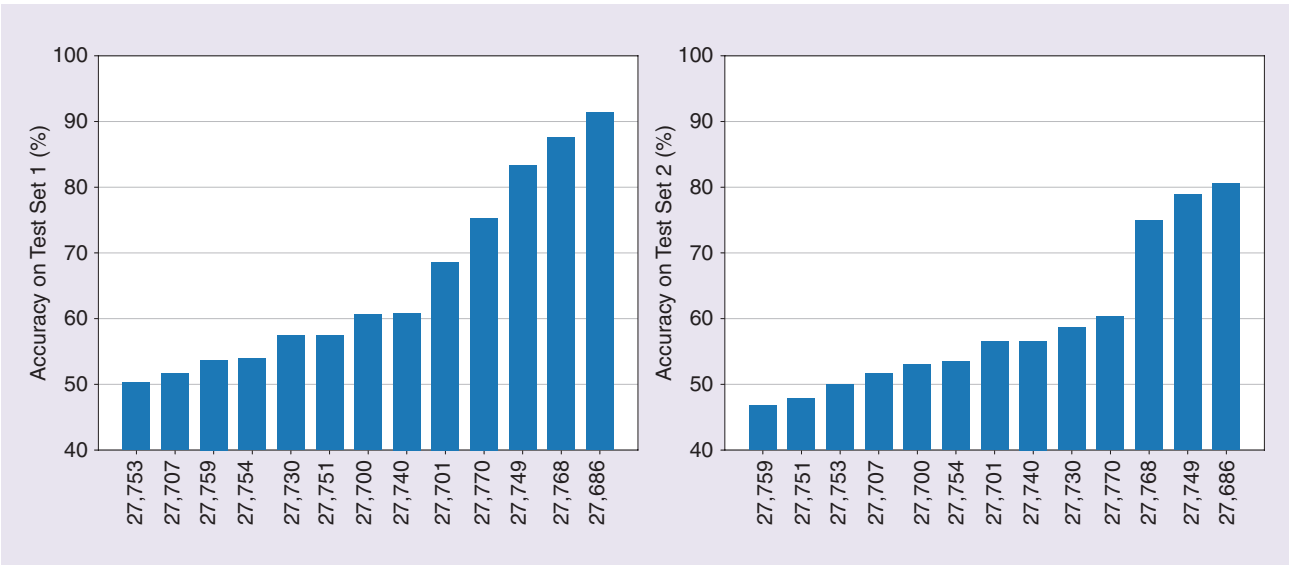

FIGURE 2. The anonymized results in terms of accuracy of the 13 teams on the two open competition datasets.

freedom for the predicting model and also to include an extra class for unknown models.

To properly capture the forensic traces that distinguish pristine images from generated ones, the networks considered multiple inputs, not just the RGB image. Indeed, it is well known that generators fail to accurately reproduce the natural correlation among color bands [20] and also that the upsampling operation routinely performed in most generative models gives rise to distinctive spectral peaks in the Fourier domain [21]. Therefore, some solutions considered as input the image represented in different color spaces, i.e., HSV and YCbCr, or computed the co-occurrence matrices on the color channels. Moreover, to exploit frequency-based features, two-stream networks have been adopted, using features extracted from the Fourier analysis in the second stream. A two-branch network was also used to work both on local and global features, which were fused by means of an attention module as done in [22]. In general, attention mechanisms have been included in several solutions. Likewise, the ensembling of multiple networks was largely used to increase diversity and boost performance. Different aggregation strategies have been pursued with the aim to generalize to unseen models and favor decisions toward the real image class, as proposed in [23].

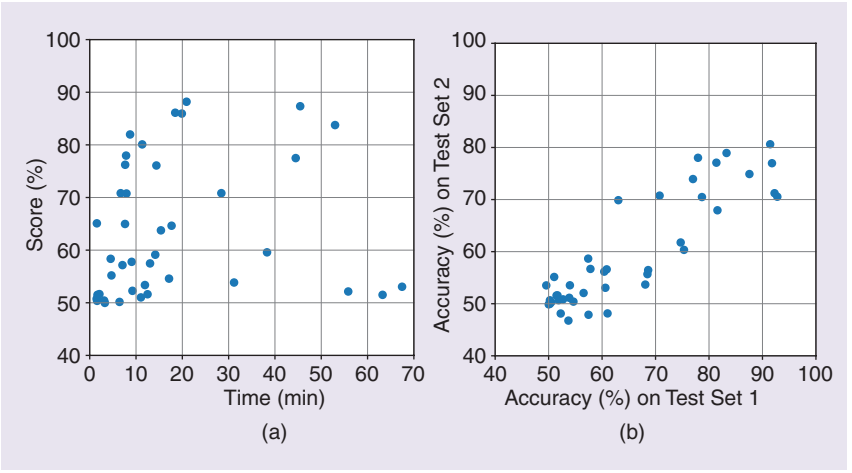

FIGURE 3. The results of all of the submitted algorithms: (a) score versus time and (b) accuracy on test set 1 versus accuracy on test set 2.

The majority of the teams trained their networks on the data made available for the challenge; however, some of them increased this dataset by generating additional synthetic images using new generative models, such as other architectures based on GANs and new ones based on diffusion models. Of course, including more generators during training helped to improve the performance, even if some approaches were able to obtain good generalization ability even adding a few more models. In addition, augmentation was always carried out to increase diversity and improve generalization. Beyond standard operations, like image flipping, cropping, resizing, and rotation, most teams used augmentation based on Gaussian blurring and JPEG compression, found to be especially helpful in the literature [24], but also changes of saturation, contrast, and brightness, as well as CutMix and random cutout.

Finalists

The final phase of the 2022 IEEE VIP Cup took place at ICIP in Bordeaux, on 16 October 2022. Figure 6 shows the members of the winning team while receiving the award. In the following, we describe the three finalist teams listed according to their final ranking: FAU Erlangen-Nürnberg (first place), Megatron (second place), and Sherlock

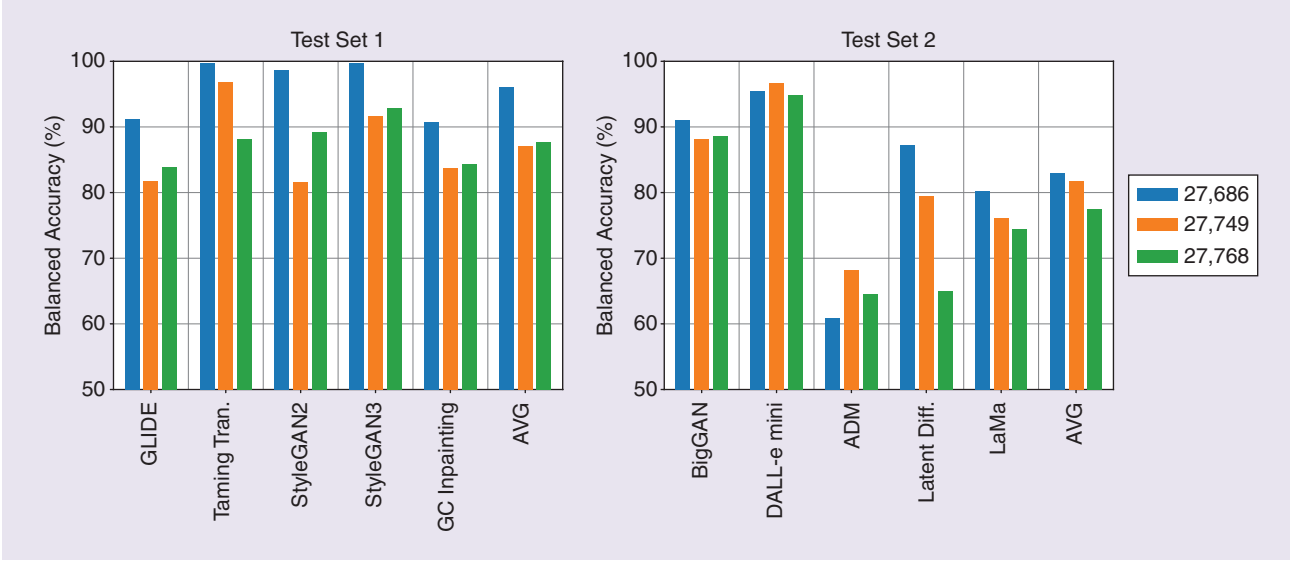

FIGURE 4. The balanced accuracy of the three best performing methods on images from test set 1 and test set 2.

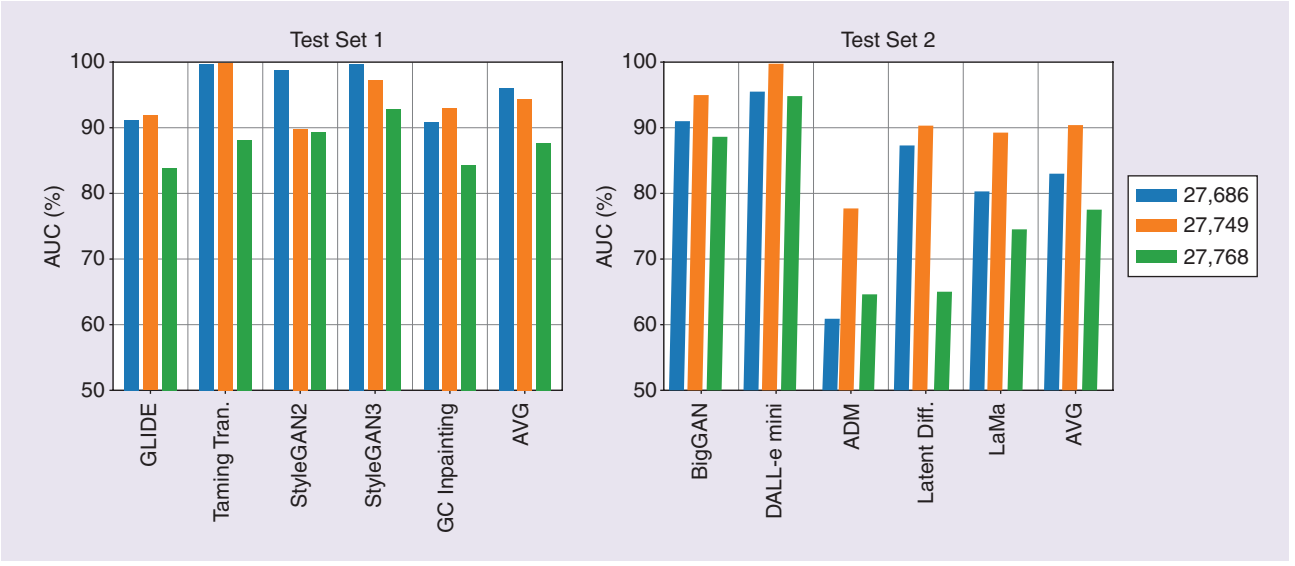

FIGURE 5. The area under the receiver operating characteristic curve (AUC) of the three best performing methods on images from test set 1 and test set 2.

FIGURE 6. The winning team (FAU Erlangen-Nürnberg) during the award ceremony at ICIP 2022 in Bordeaux.

(third place). We will also present some details on their technical approach.

FAU Erlangen-Nürnberg

- *Affiliation*: Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany
- *Supervisor*: Christian Riess
- *Tutor*: Anatol Maier
- *Students*: Vinzenz Dewor, Luca Beetz, ChangGeng Drewes, and Tobias Gessler
- *Technical approach*: an ensemble of vision transformers pretrained on Imagenet-21k and fine-tuned on a large dataset of 400,000 images. To extract generalizable features, a procedure based on weighted random sampling was adopted during training aimed at balancing the data distribution. Models included during training were the five known techniques StyleGAN2 [11], StyleGAN3 [12], GLIDE [5], Taming Transformers [10], and inpainted images with Gated Convolution [13]. In addition, images generated using DALL·E [25] and VQGAN [10] were used.

Megatron

- *Affiliation*: Bangladesh University of Engineering and Technology, Bangladesh
- *Supervisor*: Shaikh Anowarul Fattah
- *Students*: Md Awsafur Rahman, Bishmoy Paul, Najibul Haque Sarker, and Zaber Ibn Abdul Hakim
- *Technical approach*: a multiclass classification scheme and an ensemble of convolutional neural networks and transformer-based architectures. An extra class was introduced to detect synthetic images coming from unknown models. Knowledge distillation and test time augmentation were also included in the proposed solution. The training set included, beyond the five known techniques, additional images coming from the following generators: ProGAN [26], ProjectedGAN [27], CycleGAN [28], DDPM [29], Diffusion-GAN [30], Stable Diffusion [31], Denoising Diffusion GAN [32], and GauGAN [33].

Sherlock

- *Affiliation*: Bangladesh University of Engineering and Technology, Bangladesh
- *Supervisor*: Mohammad Ariful Haque
- *Students*: Fazle Rabbi, Asif Quadir, Indrojit Sarkar, Shahriar Kabir Nahin, Sawradip Saha, and Sanjay Acharjee
- *Technical approach*: a two-branch convolutional neural network that took as input features extracted in the spatial and in the Fourier domain. The adopted architectures were EfficientNet-b7 and MobileNet-v3. In addition, strong augmentation was performed, which included also CutMix beyond standard operations. During training, only the five known generation techniques were considered.

Conclusions

This article describes the 2022 VIP Cup that took place last October at ICIP. The aim of the competition was to foster research on the detection of synthetic images, in particular, focusing on images generated using the recent diffusion models [7], [8], [15], [34]. These architectures have shown an impressive ability to generate images guided by textual descriptions or pilot sketches, and there is very limited work on their detection [35], [36], [37]. Below, we highlight the main take-home messages that emerged from the technical solutions developed in this competition:

- The best-performing models are pretrained very deep networks that rely on a large dataset of real and synthetic images coming from several different generators. Indeed, increasing diversity during training was a key aspect of the best approaches.
- Augmentation represents a fundamental step to make the model more robust to post-processing operations and make it work in realistic scenarios.
- Generalization is still a main issue in synthetic image detection. In particular, it has been observed that one main problem is how to set the correct threshold in the more challenging scenario of unseen generators during training.
- The detection task can benefit of the attribution, which aims at identifying the model that was used for synthetic generation.

We believe that the availability of the dataset (https://github.com/grip-unina/DMimageDetection) created during the challenge can stimulate the research on synthetic image detection and motivate other researchers to work in this interesting field. The advancements in generative AI make the distinction between real and fake very thin, and it is very important to push the community to continuously search for effective solutions [38]. In particular, the VIP Cup has shown the need to develop models that can be used in the wild to detect synthetic images generated by new architectures, such as the recent diffusion models. In this respect, it is important to design explainable methods that can highlight which are the forensic artifacts that the detector is exploiting [39]. We hope that more and more methods will be published in the research community and will be inspired by the challenge proposed in the 2022 IEEE VIP Cup at ICIP.

In particular, the VIP Cup has shown the need to develop models that can be used in the wild to detect synthetic images generated by new architectures, such as the recent diffusion models.

Acknowledgment

The organizers express their gratitude to all participating teams, to the local organizers at ICIP 2022 for hosting the VIP Cup, and to the IEEE Signal Processing Society Membership Board for the continuous support. Special thanks go to Riccardo Corvi and Raffaele Mazza from University Federico II of Naples, who helped to build the datasets. The authors also acknowledge the projects that support this research: DISCOVER within the SemaFor program funded by DARPA under Agreement FA8750-20-2-1004; Horizon Europe vera.ai funded by the European Union, Grant Agreement 101070093; a TUM-IAS Hans Fischer Senior Fellowship; and PREMIER, funded by the Italian Ministry of Education, University, and Research within the PRIN 2017 program. This work is also funded by FCT/MCTES through national funds and when applicable cofunded EU funds under Projects UIDB/50008/2020 and LA/P/0109/2020.

Authors

Davide Cozzolino (davide.cozzolino@unina.it) is an assistant professor with the Department of Electrical Engineering and Information Technology, University Federico II, 80125 Naples, Italy. He was cochair of the IEEE CVPR Workshop on Media Forensics in 2020. He was part of the teams that won the 2013 IEEE Image Forensics Challenge (both detection and localization) and the 2018 IEEE Signal Processing Cup on camera model identification. His research interests include image processing and deep learning, with main contributions in multimedia forensics. He is a Member of IEEE.

Koki Nagano (knagano@nvidia.com) is a senior research scientist at NVIDIA Research, Santa Clara, CA 95051 USA. He works at the intersection of graphics and AI, and his research focuses on realistic digital human synthesis and trustworthy visual computing including the detection and prevention of visual misinformation. He is a Member of IEEE.

Lucas Thomaz (lucas.thomaz@co.it.pt) is a researcher at Instituto de Telecomunicações, 2411-901 Leiria, Portugal, and an associate professor in the School of Technology and Management, Polytechnic of Leiria, Leiria 2411-901, Portugal. He is a member of the Student Services Committee of the IEEE Signal Processing Society, supporting the VIP Cup and the IEEE Signal Processing Cup, and the chair of the Engagement and Career Training Subcommittee. He

is a consulting associate editor for *IEEE Open Journal of Signal Processing*. He is a Member of IEEE.

Angshul Majumdar (angshul@iiitd.ac.in) received his Ph.D. from the University of British Columbia. He is a professor at Indraprastha Institute of Information Technology, New Delhi 110020, India. He has been with the institute since 2012. He is currently the director of the Student Services Committee of the IEEE Signal Processing Society. He has previously served the Society as chair of the Chapter's Committee (2016–2018), chair of the Education Committee (2019), and member-at-large of the Education Board (2020). He is an associate editor for *IEEE Open Journal of Signal Processing* and Elsevier's *Neurocomputing*. In the past, he was an associate editor for *IEEE Transactions on Circuits and Systems for Video Technology*. He is a Senior Member of IEEE.

Luisa Verdoliva (verdoliv@unina.it) is a professor with the Department of Electrical Engineering and Information Technology, University Federico II, 80125 Naples, Italy. She was an associate editor for *IEEE Transactions on Information Forensics and Security* (2017–2022) and is currently deputy editor in chief for the same journal and senior area editor for *IEEE Signal Processing Letters*. She is the recipient of a Google Faculty Research Award for Machine Perception (2018) and a TUM-IAS Hans Fischer Senior Fellowship (2020–2024). She was chair of the IFS TC (2021–2022). Her scientific interests are in the field of image and video processing, with main contributions in the area of multimedia forensics. She is a Fellow of IEEE.

References

[1] A. Mahdawi, "Nonconsensual deepfake porn is an emergency that is ruining lives," *Guardian*, Apr. 2023. [Online]. Available: https://www.theguardian.com/commentisfree/2023/apr/01/ai-deepfake-porn-fake-images

[2] J. Vincent, "After deepfakes go viral, AI image generator Midjourney stops free trials citing 'abuse'," *Verge*, Mar. 2023. [Online]. Available: https://www.theverge.com/2023/3/30/23662940/deepfake-viral-ai-misinformation-midjourney-stops-free-trials

[3] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 5, pp. 910–932, Aug. 2020, doi: 10.1109/JSTSP.2020.3002101.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 4396–4405, doi: 10.1109/CVPR.2019.00453.

[5] A. Q. Nichol et al., "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 16,784–16,804.

[6] B. Dayma et al. "DALL-E mini." GitHub. Accessed: Dec. 14, 2022. [Online]. Available: https://github.com/borisdayma/dalle-mini

[7] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2022, pp. 10,684–10,695.

[8] Y. Balaji et al., "eDiff-I: Text-to-image diffusion models with ensemble of expert denoisers," 2022, *arXiv:2211.01324*.

[9] 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 *Proc. IEEE Int. Conf. Multimedia Expo*, 2021, pp. 1–6, doi: 10.1109/ICME51207.2021.9428429.

[10] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2021, pp. 12,873–12,883, doi: 10.1109/CVPR46437.2021.01268.

[11] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 8110–8119.

[12] T. Karras et al., "Alias-free generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 852–863.

[13] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vision*, 2019, pp. 4471–4480, doi: 10.1109/ICCV.2019.00457.

[14] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019, pp. 1–11.

[15] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 8780–8794.

[16] R. Suvorov et al., "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vision*, 2022, pp. 3172–3182, doi: 10.1109/WACV51458.2022.00323.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit.*, 2009, pp. 248–255, doi: 10.1109/CVPR.2009.5206848.

[18] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vision*, 2014, pp. 740–755.

[19] F. Yu, A. Seff, Y. Zhang, S. Song, T. Funkhouser, and J. Xiao, "LSUN: Construction of a large-scale image dataset using deep learning with humans in the loop," 2015, *arXiv:1506.03365*.

[20] H. Li, B. Li, S. Tan, and J. Huang, "Identification of deep network generated images using disparities in color components," *Signal Process.*, vol. 174, Sep. 2020, Art. no. 107616, doi: 10.1016/j.sigpro.2020.107616.

[21] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in GAN fake images," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2019, pp. 1–6, doi: 10.1109/WIFS47025.2019.9035107.

[22] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu, "Fusing global and local features for generalized AI-synthesized image detection," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 3465–3469, doi: 10.1109/ICIP46576.2022.9897820.

[23] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro, "Detecting GAN-generated images by orthogonal training of multiple CNNs," in *Proc. IEEE Int. Conf. Image Process.*, 2022, pp. 3091–3095, doi: 10.1109/ICIP46576.2022.9897310.

[24] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. Efros, "CNN-generated images are surprisingly easy to spot… for now," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2020, pp. 8692–8701, doi: 10.1109/CVPR42600.2020.00872.

[25] A. Ramesh et al., "Zero-shot text-to-image generation," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.

[26] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018, pp. 1–12.

[27] A. Sauer, K. Chitta, J. Müller, and A. Geiger, "Projected GANs converge faster," in *Proc. Adv. Neural Inf. Process. Syst.*, 2021, pp. 1–13.

[28] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vision*, 2017, pp. 2242–2251, doi: 10.1109/ICCV.2017.244.

[29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 6840–6851.

[30] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-GAN: Training GANs with diffusion," in *Proc. Int. Conf. Learn. Representations*, 2023, pp. 1–13.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. "Stable diffusion." GitHub. Accessed: Dec. 14, 2022. [Online]. Available: https://github.com/CompVis/stable-diffusion

[32] Z. Xiao, K. Kreis, and A. Vahdat, "Tackling the generative learning trilemma with denoising diffusion GANs," in *Proc. Int. Conf. Learn. Representations*, 2022, pp. 1–15.

[33] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2019, pp. 2332–2341.

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

[35] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2023, pp. 1–5.

[36] Z. Sha, Z. Li, N. Yu, and Y. Zhang, "DE-FAKE: Detection and attribution of fake images generated by text-to-image diffusion models," 2022, *arXiv:2210.06998*.

[37] J. Ricker, S. Damm, T. Holz, and A. Fischer, "Towards the detection of diffusion model deepfakes," 2022, *arXiv:2210.14571*.

[38] M. Barni et al., "Information forensics and security: A quarter century long journey," *IEEE Signal Process. Mag.*, vol. 40, no. 5, pp. 67-79, Jul. 2023, doi: 0.1109/MSP.2023.3275319.

[39] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: From generative adversarial networks to diffusion models," in *Proc. IEEE Comput. Vision Pattern Recognit. Workshops*, 2023, pp. 973–982.

SP