Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 246 (2024) 4160–4168

Procedia Computer Science

www.elsevier.com/locate/procedia

28th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES 2024)

Facial Segmentation in Deepfake Classification: a Transfer Learning Approach

Marco Tanfoni[a,*], Elia Giuseppe Ceroni[a,*], Niccolò Pancino[a], Monica Bianchini[a], Marco Maggini[a]

[a]*University of Siena - Department of Information Engineering and Mathematics; Via Roma, 56, 53100 - Siena, Italy*

Abstract

Artificial Intelligence (AI)–generated images represent a significant threat in various fields, such as security, privacy, media forensics and content moderation. In this paper, a novel approach for the detection of StyleGAN2–generated human faces is presented, leveraging a Transfer Learning strategy to improve the classification performance of the models. A modified version of the state–of–the–art semantic segmentation model DeepLabV3+, using either a ResNet50 or a MobileNetV3 Large as feature extraction backbones, is used to create both a face segmentation model and the synthetic image detector. To achieve this goal, the models are at first trained for face segmentation in a multi–class classification task on a widely used semantic segmentation dataset, achieving remarkable results for both configurations. Then, the pre–trained models are retrained on a collection of real and generated images, gathered from different sources to solve a binary classification task, namely to detect synthetic (i.e. generated) images, thus carrying out two different transfer learning strategies. The results indicate that this targeted methodology significantly improves the detection rates compared to analyzing the face as a whole, and underlines the importance of advanced image recognition technologies when tackling the challenge of detecting generated faces.

Keywords: Fake detection; Image Authentication; Synthetic Image Detection; Image segmentation; Transfer learning; DeepLabV3+; MobileNetV3; ResNet; Digital Forensics; Machine Learning; Computer Vision

1. Introduction

With the emergence of many state–of–the–art generative models for human face generation, such as NVIDIA's StyleGAN [1] or Stable Diffusion [2], a vast number of synthetic images have proliferated over the internet, particularly in social media. This, along with the fact that many of the fake images are nearly indistinguishable to the human eye from the real ones, poses a concerning threat to trustworthiness in online environments. The ability to recognize

* Equal contribution and corresponding authors
E-mail address: marco.tanfoni@student.unisi.it, elia.ceroni@student.unisi.it

1877-0509
10.1016/j.procs.2024.09.255

generated images from real photos has thus become a critical concern in various fields such as media forensics, security and content moderation. Despite having been addressed in the literature [3, 4] with positive results, there is still room to expand on this topic in terms of analyzing facial features singularly to improve classification results. In particular, most of the models in literature are based on Convolutional Neural Networks (CNNs, [5]), a class of deep learning models specifically designed for analyzing image and video data [6, 7]: CNNs are able to learn spatial hierarchies of features by means of multiple convolutional, pooling, and fully connected layers. From the seminal works in [8, 9], CNNs have revolutionized various fields, including computer vision [10, 11, 12], medical image analysis [13, 14, 15, 16, 17], and autonomous driving [18, 19, 20, 21]. In addition to image classification, CNNs have been successfully applied to different tasks such as object detection, semantic segmentation [22, 23, 24], and image generation [25, 26]. In this context, the ability to accurately separate real and AI–generated images became very important after the quick advancement of generative technologies, like the ones based on Generative Adversarial Networks (GANs [25]): these models are remarkably useful in many tasks such as synthetic data generation, in which they facilitated the spread of hyper–realistic facial images, not easily distinguishable to the untrained human eye from genuine photos. Among the most influential GAN–based models, which have significantly contributed to this trend, are those in the StyleGAN [1, 27, 28] and BigGAN [29] families; each of these pushed the boundaries of image quality and resolution. StyleGAN2 model [27], for instance, refined the architecture to eliminate some artifacts and further enhance image quality. BigGAN [29], on the other hand, is known for generating high–fidelity images at substantially larger scales than previous GANs. These advancements have made it increasingly difficult to distinguish computer–generated faces from real human images, intensifying the need for robust detection mechanisms. There have been various recent studies contributing with multiple methodologies to address the task of detecting such images. For instance, [30] introduces a specialized dataset called OpenForensics, designed with face–wise rich annotations useful for training models capable of performing multi–face forgery detection and segmentation in in–the–wild scenes, providing potentials for research in both deepfake prevention and general human face detection. Furthermore, [31] introduced other datasets and methods which incorporate facial landmarks to locate edited facial components, thus promoting more granular and interpretable face verification technologies. Similarly, [32] approaches deepfake detection as a fine–grained classification problem, focusing on discriminative facial subtle features to enhance detection across various datasets and image manipulations. These methods significantly improve robustness and generalization in deepfake data detection. In [33], the authors address the challenge of GAN–generated faces with a novel architecture called Gram–Net, based on global image texture features for enhanced detection, thus highlighting the importance of texture as a reliable indicator of image authenticity. Moreover, [34] introduced DETER, an approach that leverages the analysis of layer–by–layer neuron activation patterns to detect AI–generated images, proving itself robust against a variety of fake faces generated by GANs, as well as against common adversarial attacks, thus setting a new baseline for fake detection technologies.

This study proposes to tackle this problem by focusing independently on different parts of the generated face, so as to obtain a classification based on human–understandable extracted features. To achieve this goal, a multiple step approach, based on DeepLabV3+ [22] model, is introduced: in a preliminary phase, the whole model is trained from scratch (i.e., DeepLabV3+ parameters are re–initialized) directly on the binary classification task, using DeepLabV3+ as a backbone purely as a feature extraction model. To asses the classification capabilities of the model, a transfer learning [35, 36] approach was exploited as well, to improve performance in terms of learning times and evaluation metrics. The implementation details and code are available online in a public GitHub repository[1]. As a matter of fact, thanks to its state–of–the–art semantic segmentation capabilities, a transfer learning approach on DeepLabV3+ was exploited in the study: hence, it was used to locate various salient parts of the face, e.g., eyes, ears or nose, in order to have a latent useful representation of the facial areas. In this context, the model was pre–trained on a specific semantic segmentation dataset in a multi–class classification task, and then exploited in the binary classification task, in which two transfer learning settings were explored, by re–training the whole model as well as keeping most of the layers non–trainable with the exception of the head layers.

The rest of the paper is structured as follows: Section 2 describes the employed methodologies, including the dataset description for both the image segmentation and binary classification tasks, as well as the model architecture and learning procedure. Section 3 shows the experimental setup chosen for the study, with specific reference to data

[1] https://github.com/mtanf/gng

setting, model selection, and evaluation. Section 4 presents the results and discusses the performance of the model. Finally, Section 5 draws some conclusions on current research and explains future perspectives.

2. Materials and Methods

In this section, the data and the experimental methodology used in this work are described: Section 2.1 presents an overview of DeepLabV3+ model, used in all the learning settings, while Section 2.2 proposes a precise description of the datasets for both the multi–class and the binary classification tasks.

2.1. DeepLabV3+

DeepLabV3+ is the latest model belonging to the DeepLab [23] family, a widely used class of convolutional neural network (CNN) models for semantic segmentation, developed by Google Research. A key characteristic of this family of models is the usage of Atrous Spatial Pyramid Pooling (ASPP) [23]. This technique, introduced for DeepLabv2, combines Atrous convolutions and Spatial Pyramid Pooling operations to allow the capture of meaningful information at multiple scales, making it particularly suited for semantic segmentation tasks. In atrous convolution, the kernels of the layer are expanded by introducing gaps between filter layers. This leads to an increase of the receptive field of the neurons in the layer, without increasing the number of parameters, also capturing features at multiple scales. Spatial Pyramid Pooling (SPP) [37] is a pooling strategy capable of generating fixed length representations of images of varying size. In SPP, the input feature maps are divided into sub–regions at different scales. Each sub–region is then pooled independently for feature extraction and, lastly, the pooled features are concatenated. The SPP layer allows the network to efficiently capture both local and global spatial information and be resistant to variations in object positioning and size. DeepLabv3+ [22] extends the DeepLabv3 model [23] by using an Encoder–Decoder structure: the backbone encoder model extracts rich semantic information by applying dilated convolution at multiple scales, which is then used by the decoder to determine detailed boundaries for objects in the image.

2.2. Dataset Description

The complete version of Mut1ny [38] was exploited to train the DeepLabV3+ segmentation module. The dataset contains 70621 images labelled at the pixel level. It includes facial images spanning various ethnicities, ages, and genders, thus making it a balanced dataset widely used in literature [39, 40, 41]. Additionally, it features a diverse array of facial poses and camera angles, ensuring comprehensive coverage from all angles. The 14 head/face segmentation classes, each representing different facial features and attributes, are reported in Fig. 1. In the generated (i.e. fake) images' detection task, an ensemble of datasets was constructed, by retrieving real images from publicly available datasets, and by generating fake facial images using state–of–the–art ML models. Some of the datasets included in the ArtiFact [42] — a large collection of real and synthetic images, including various categories like people, animals, places, vehicles, and more — were used to retrieve real (i.e. not generated) images: in particular, images were extracted from CelebA–HQ [43] and FF–HQ dataset. The final dataset used for the learning procedure comprises 100,000 real and 200,000 synthetic facial images (the real images include 70,000 from the FF–HQ dataset and 30,000 from CelebA–HQ dataset). On the other hand, a set of generative models, including Diffusion GANs [44], Stable Diffusion models, StyleGAN1, StyleGAN2 [27], StyleGAN3 [28], Palette, and ProjectedGAN, [45] were used for the generation of the positive samples.

3. Experimental Setup

This section defines the experimental setup chosen for the study, with specific reference to data preprocessing, model selection, and evaluation. In the semantic segmentation learning procedure, augmentation techniques were applied on training data. In particular, images were randomly rotated within a range of ±15 degrees to simulate variations in head pose, as well as translated horizontally and vertically up to 10% of their total width and height, respectively, to account for misalignment. A shear transformation of up to 0.2 radians was applied, altering the geometry of images to simulate a change in perspective. Moreover, random zooming of images by up to 20% to mimic distance variations

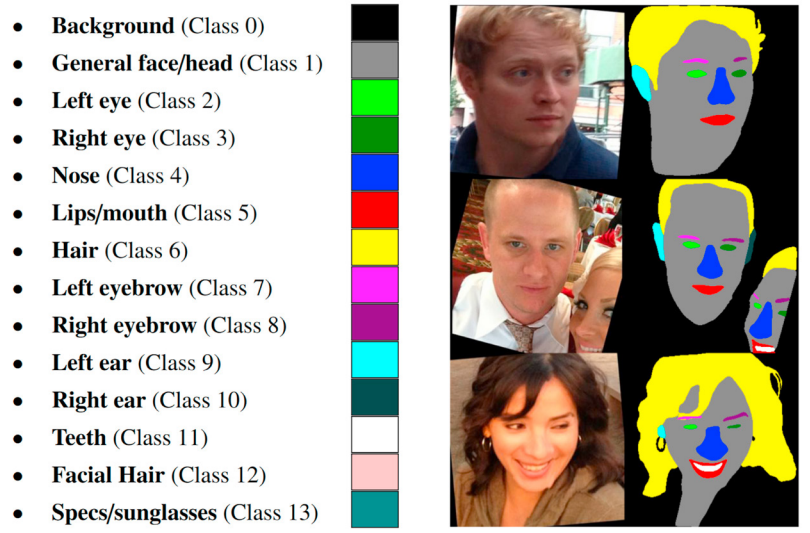

Fig. 1: Examples of segmentation masks in Mut1ny dataset. Each color in the segmentation mask corresponds to a specific facial element, as denoted in the legend on the left.

from the camera was applied, and then images were randomly flipped horizontally to simulate different orientations. Finally, a "nearest" method was used to fill in new pixels created by rotations or width/height shifts, preserving the local pixel similarity. The proposed architecture is depicted in Fig. 2. In the this version of the model, a ResNet50 [47] and a MobileNetV3Large [48] were used as feature extractors, in place of the original Xception [49] backbone. Values of dilation rates of 1, 4, 8, and 16 were used in the ASPP module, to obtain a good balance between capturing large context and fine details. Moreover, the ASPP layer was modified to increase the kernel size (from the standard 3×3 dimension), along with the dilation rate, obtaining kernels of shape 3, 5, 7, and 11, in order to improve the gathering of contextual information at this stage of the network.

To train the segmentation model, a Sparse Categorical Cross–Entropy loss was used, while the deepfake classifier network was trained with a Binary Cross–Entropy loss. For both problems, a set of class weights have been calculated, based on the numerosity of each class in the dataset, to handle class inbalance issues. This is especially significant in the semantic segmentation task, due to the presence of classes (such as teeth and glasses) which encompass a small number of pixels. Adam optimizer was used in the learning procedure, with an initial learning rate of 0.001.

Moreover, a transfer learning (TL) [35, 36] strategy was also implemented. The weights of the network trained on the segmentation task are used as starting point for the training of the deepfake detector network. This was done since the face landmark segmentation task should have some degree of correlation with the deepfake classification problem, since current generation GAN models are not always consistent with regard to the coherence of facial feature position and shapes [50, 51]. The primary benefits of the TL approach should be reduced training times, less samples required in the learning procedure, and possibly, improved performance. These advantages are particularly significant for deep networks, since they usually need huge amounts of data and extensive training periods to achieve optimal performance.

Evaluation metrics in the considered generated images' detection task are the ones usually employed for binary classification problems, in which the performance of the model is measured in terms of accuracy, precision, recall and F1–Score, as defined in Eq. (1).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \qquad \text{Precision} = \frac{TP}{TP + FP} \qquad (1)$$
$$\text{Recall} = \frac{TP}{TP + FN} \qquad \text{F1–Score} = \frac{2TP}{2TP + FP + FN}$$

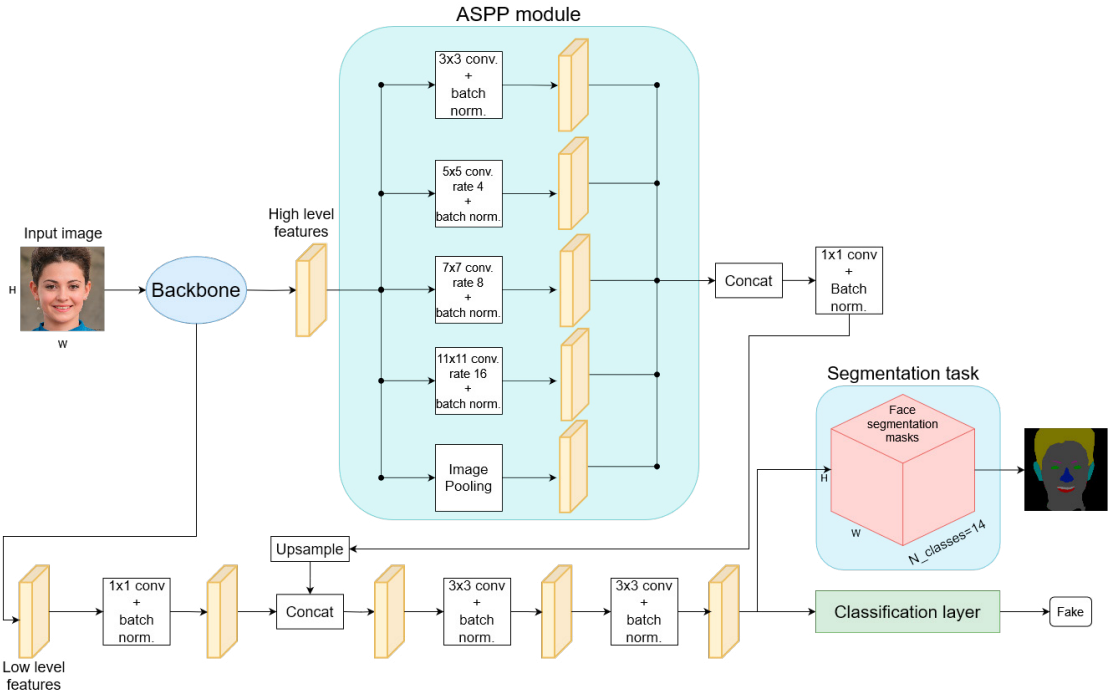

Fig. 2: Network architecture. This model is used for both the semantic segmentation and fake detection tasks. In both the face segmentation task and the baseline fake detection experiment, the network is fully trainable and initialized with the ImageNet [46] parameters. On the other hand, in the transfer learning experiments for the fake detection problem, the models are initialized with the parameters from the face segmentation task, and then two strategies — depending on the retrainability of the backbone model together with the head layers — are carried out.

In particular, the different outcomes of predictions made by a generic model compared to the actual ground truth are described by using true positives, true negatives, false positives, and false negatives. A true positive (TP) occurs when the model accurately identifies an instance as belonging to the positive class, whereas a true negative (TN) occurs when it correctly identifies an instance as belonging to the negative class. In contrast, a false positive (FP) occurs when the model erroneously classifies a negative instance as positive, and a false negative (FN) occurs when it incorrectly labels a positive instance as negative. These distinctions are significant in the evaluation procedure of classification models, giving insight into their accuracy, reliability, and information for potential improvement. It is worth noting that the F1–Score can be interpreted as a weighted average of the precision and recall, thus providing a more balanced measure in cases of unbalanced datasets.

4. Results and Discussion

Table 1 reports a summary of the DeepLabV3+ based face segmentators on the Mut1ny test set images, for both ResNet50 and MobileNetV3 Large backbone setups. Both feature extraction backbones deliver satisfactory results and, indeed, the performance gap between the ResNet50 and the MobileNetV3 backbones is relatively small, despite the latter having approximately one–third of the parameters compared to the former.

Some results of the application of the trained face segmentation modules applied to both real and generated images are depicted in Fig. 3. It is remarkable how, despite different light and focus conditions, the model proves to be precise on the segmentation task, even for partially occluded faces like the first one. The robustness of the model underscores its adaptability to multiple real–world scenarios where conditions are not always ideal, thus creating opportunities for deployment in fields such as public security systems and mobile applications, also considering the lightweight nature of the MobileNetV3 Large backbone.

Table 2 reports the results in the binary fake detection classification task for all the considered settings.

Table 1: Summary of DeepLabV3+ segmentation models performance. The feature extraction backbone architecture used is reported, together with the total number of trainable parameters and the pixel accuracy of the trained model on the test section of the complete Mut1ny dataset.

Backbone	Number of parameters	Accuracy
ResNet50	15,411,966	**0.9539**
MobileNetV3 Large	**5,815,646**	0.9484

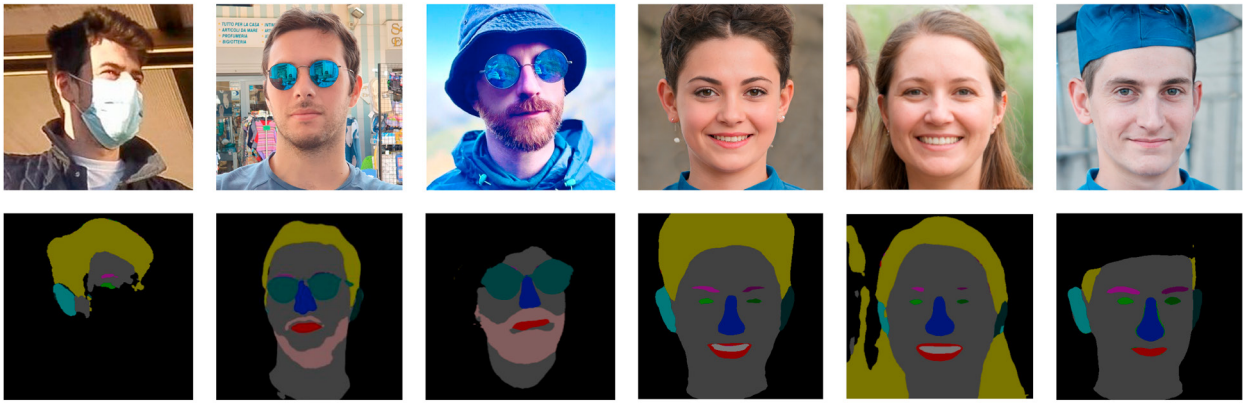

Fig. 3: From the left: input images and segmentation masks obtained from the trained segmentation model, on three real images and three StyleGAN2–generated images, retrieved from www.thispersondoesnotexist.com.

Table 2: Fake detection performance on StyleGAN2–generated images in the test set. The column TL refers to the transfer learning strategy used in the learning procedure: "—" means that no transfer learning is applied, *Last* means that only the backbone feature extractor and final layers are trainable, while the rest of the sub–model is set as non–trainable, *All* means that all the model parameters are re–trainable, using the pre–trained parameters only for the initialization of the whole model. Both transfer learning strategies improve the performance of the classifier, for both the feature extractor backbones, showing the feasibility of the approach.

Backbone	TL	Accuracy	Precision	Recall	F1
ResNet50	—	0.8704	0.8751	0.9397	0.9062
	Last	0.9176	0.9376	0.9387	0.9382
	All	**0.9342**	**0.9509**	**0.9504**	**0.9507**
MobileNetV3 Large	—	0.8690	0.9073	0.8950	0.9011
	Last	0.8840	**0.9842**	0.8395	0.9061
	All	**0.9504**	0.9623	**0.9634**	**0.9628**

The results indicate that pre–training the deepfake classifier on the segmentation task can enhance the model's ability to adapt to the subtle differences between synthetic and real facial features. This is in line with the expectations from the literature on transfer learning, and underlines the effectiveness of this approach for improving detection rates. Furthermore, the low–level understanding of facial features granted by the segmentation model allows for more detailed analyses and interpretation of model–specific incongruities in generated images. This could also help paving the way for employment in the image–forgery field, that is, manipulating real images to deceive viewers, typically associated with malicious activities, with potential consequences in fields like journalism, personal privacy and legal proceedings. The adaptability of the model across the considered different architectures also demonstrates the versatility of the approach, suggesting that it could also be applied to other setups. In particular, the transfer learning strategy had a remarkable effect for both the backbone models over the non–transfer learning cases, with an improvement in performance of about 8% for the MobileNet submodule and of about 7% in the ResNet submodule, when the pre–

trained models are used only as parameters initializers. That is, starting from a favorable situation in the parameters initialization, the models can adapt the hidden representation of the samples to the task they are trained for. This result is in accordance with known literature [10], that highlights how large models, as the ones used in the study as segmentation submodule, benefit from transfer learning. With respect to the final classification performance, the results are in line with the literature regarding StyleGAN2–generated images detection, as reported in [4], even considering the heterogeneity of approaches and evaluation methods there reported. However, it is worth noting that, to the authors' knowledge, using transfer learning from segmentation models to generated images detection is a novel approach, which could possibly enhance current methodologies, so as to obtain ever higher detection rates. Additionally, the original Xception backbone featured in the DeepLabV3+ architecture has been tested. Although this provided slightly better segmentation performance, this improvement came at the cost of about 12 times the number of parameters (~75M *vs.* ~6M with the MobileNet backbone), thus greatly extending training time and reducing practical feasibility for many applications.

Interestingly enough, although MobileNetV3 Large segmentation capabilities are not as remarkable as the ones of the ResNet50 submodule, it performed overall better in the fake detection task: a key aspect of the two aformentioned models is the difference in the number of trainable parameters, since ResNet50 has more than three times the number of parameters compared to MobilenetV3, which may favour one over the other in some practical applications.

5. Conclusion

This paper introduces a novel approach for the detection of AI–generated human faces, which leverages a Transfer Learning schema to improve the classification performance of the models, integrating a pre–trained segmentation model capable of identifying various facial landmarks. Three experiments were performed, training the model from scratch, and by using two transfer learning policies in which the same architecture was pre–trained specifically for a segmentation task, and then used in the fake detection task, thus adopting either a partial or a full transfer learning approach. The latter has shown not only to improve the fake detection performance, but also to produce interesting and human interpretable insights, by deploying ad–hoc interpretability or explainability methods to provide information on the most crucial features for the classification, thus also pushing the boundaries towards more reliable artificial intelligence in sensitive areas such as digital forensics and media integrity.

Possible limitations of the proposed approach include the fact that FFHQ, which represents a large fraction of the real faces dataset used in the work is also the training dataset for StyleGAN networks, adding possible biases to the method. Moreover, the synthetic faces dataset consists of images generated only with StyleGAN2, thus leaving the generalization to other generator networks an open problem. Lastly, since the TL scheme used here can be applied to other architectures, exploring other models beyond DeepLabV3+ might yield to potentially higher performance.

It is a matter of future research to expand the dataset diversity, including other collections of both AI–generated and real human images, also obtained from the latest generative models. Furthermore, given the lightweight nature of MobileNetV3 and the relatively comparable performance with respect to the ResNet50, the DeepLabV3+ model with this backbone feature extractor could also be employed in real–time data stream and on devices with lower computational power, thus extending its applicability to other contexts, such as surveillance and content moderation. Finally, implementing explainability tools such as SHAP [52] or Grad–cam [53] to show the most important facial features, or even to guide the training procedure, will be useful to highlight weak points of a certain generator and consequently obtain better performance.

References

[1] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, 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. 10684–10695.

[3] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, J. Ortega-Garcia, Deepfakes and beyond: A survey of face manipulation and fake detection, Information Fusion 64 (2020) 131–148. doi:https://doi.org/10.1016/j.inffus.2020.06.014.

[4] X. Wang, H. Guo, S. Hu, M.-C. Chang, S. Lyu, Gan-generated faces detection: A survey and new perspectives (2023). arXiv:2202.07145.

[5] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, nature 521 (7553) (2015) 436–444.

[6] R. Yamashita, M. Nishio, R. K. G. Do, K. Togashi, Convolutional neural networks: an overview and application in radiology, Insights into imaging 9 (2018) 611–629.
[7] Z. Li, F. Liu, W. Yang, S. Peng, J. Zhou, A survey of convolutional neural networks: analysis, applications, and prospects, IEEE transactions on neural networks and learning systems 33 (12) (2021) 6999–7019.
[8] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, Backpropagation applied to handwritten zip code recognition, Neural computation 1 (4) (1989) 541–551.
[9] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.
[10] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, Advances in neural information processing systems 25 (2012) 1097–1105.
[11] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition (2014) 580–587.
[12] M. Monaci, N. Pancino, P. Andreini, S. Bonechi, P. Bongini, A. Rossi, G. Ciano, G. Giacomini, F. Scarselli, M. Bianchini, et al., Deep learning techniques for dragonfly action recognition., in: ICPRAM, 2020, pp. 562–569.
[13] E. Landi, F. Spinelli, M. Intravaia, M. Mugnaini, A. Fort, M. Bianchini, B. T. Corradini, F. Scarselli, M. Tanfoni, A mobilenet neural network model for fault diagnosis in roller bearings, in: 2023 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), 2023, pp. 01–06. doi:10.1109/I2MTC53148.2023.10176049.
[14] N. Pancino, C. Graziani, V. Lachi, M. L. Sampoli, E. Ștefănescu, M. Bianchini, G. M. Dimitri, A mixed statistical and machine learning approach for the analysis of multimodal trail making test data, Mathematics 9 (24) (2021) 3159.
[15] M. Intravaia, A. Fort, E. Landi, M. Mugnaini, M. Bianchini, B. Corradini, F. Scarselli, M. Tanfoni, F. Spinelli, et al., A mobilenet neural network model for fault diagnosis in roller bearings, in: Proceedings of IEEE I2MTC 2023 Conference, IEEE, 2023.
[16] E. Stefanescu, N. Pancino, C. Graziani, V. Lachi, M. Sampoli, G. Dimitri, A. Bargagli, D. Zanca, M. Bianchini, D. Mureșanu, et al., Blinking rate comparison between patients with chronic pain and parkinson's disease, EUROPEAN JOURNAL OF NEUROLOGY 29 (2022) 669–669.
[17] V. Russo, E. Lallo, A. Munnia, M. Spedicato, L. Messerini, R. D'Aurizio, E. G. Ceroni, G. Brunelli, A. Galvano, A. Russo, et al., Artificial intelligence predictive models of response to cytotoxic chemotherapy alone or combined to targeted therapy for metastatic colorectal cancer patients: a systematic review and meta-analysis, Cancers 14 (16) (2022) 4012.
[18] S. Lee, J. Lee, K. M. Lee, V-net: End-to-end convolutional network for object detection, Expert Systems with Applications 90 (2017) 295–304.
[19] J. Liang, N. Li, X. Sun, X. Wang, M. Liu, J. Shi, J. Huang, D.-Y. Wang, Cirl: Continuous imitation learning from human interaction with reinforcement learning in autonomous driving, IEEE Transactions on Intelligent Transportation Systems 20 (11) (2018) 4038–4052.
[20] T. Chen, S. Liu, X. Yang, J. Shen, X. Hu, G. Yang, Deepdriving: Learning affordance for direct perception in autonomous driving, Proceedings of the AAAI Conference on Artificial Intelligence 29 (1) (2015) 2722–2728.
[21] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al., End to end learning for self-driving cars, in: arXiv preprint arXiv:1604.07316, 2016.
[22] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, H. Adam, Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 801–818.
[23] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A. L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE transactions on pattern analysis and machine intelligence 40 (4) (2017) 834–848.
[24] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, Proceedings of the IEEE conference on computer vision and pattern recognition (2015) 3431–3440.
[25] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).
[26] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, CoRR abs/1511.06434 (2015).
URL https://api.semanticscholar.org/CorpusID:11758569
[27] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8110–8119.
[28] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, T. Aila, Alias-free generative adversarial networks, Advances in neural information processing systems 34 (2021) 852–863.
[29] A. Brock, J. Donahue, K. Simonyan, Large scale gan training for high fidelity natural image synthesis (2019). arXiv:1809.11096.
[30] T.-N. Le, H. H. Nguyen, J. Yamagishi, I. Echizen, Openforensics: Large-scale challenging dataset for multi-face forgery detection and segmentation in-the-wild, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 10097–10107. doi:10.1109/ICCV48922.2021.00996.
[31] K. Songsri-in, S. Zafeiriou, Complement face forensic detection and localization with faciallandmarks (2019). arXiv:1910.05455.
[32] A. V. Nadimpalli, A. Rattani, Facial forgery-based deepfake detection using fine-grained features, in: 2023 International Conference on Machine Learning and Applications (ICMLA), 2023, pp. 2174–2181. doi:10.1109/ICMLA58977.2023.00328.
[33] Z. Liu, X. Qi, P. H. Torr, Global texture enhancement for fake face detection in the wild, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 8060–8069.
[34] R. Wang, F. Juefei-Xu, L. Ma, X. Xie, Y. Huang, J. Wang, Y. Liu, Fakespotter: a simple yet robust baseline for spotting ai-synthesized fake faces, in: Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI'20, 2021.
[35] L. Torrey, J. Shavlik, Transfer learning, in: Handbook of research on machine learning applications and trends: algorithms, methods, and techniques, IGI global, 2010, pp. 242–264.
[36] K. Weiss, T. M. Khoshgoftaar, D. Wang, A survey of transfer learning, Journal of Big data 3 (2016) 1–40.

[37] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, IEEE transactions on pattern analysis and machine intelligence 37 (9) (2015) 1904–1916.
[38] Mut1ny, Face/head segmentation dataset commercial purpose edition (March 2024).
[39] A. Hassani, Z. E. Shair, R. Ud Duala Refat, H. Malik, Distilling facial knowledge with teacher-tasks: Semantic-segmentation-features for pose-invariant face-recognition, in: 2022 IEEE International Conference on Image Processing (ICIP), 2022, pp. 741–745. doi:10.1109/ICIP46576.2022.9897793.
[40] M. Reimann, M. Klingbeil, S. Pasewaldt, A. Semmo, M. Trapp, J. Döllner, Locally controllable neural style transfer on mobile devices, The Visual Computer 35 (11) (2019) 1531–1547. doi:10.1007/s00371-019-01654-1.
URL https://doi.org/10.1007/s00371-019-01654-1
[41] E. Khoshnevisan, H. Hassanpour, M. M. AlyanNezhadi, Face recognition based on general structure and angular face elements, Multimedia Tools and Applications (2024). doi:10.1007/s11042-024-18897-3.
URL https://doi.org/10.1007/s11042-024-18897-3
[42] M. A. Rahman, B. Paul, N. H. Sarker, Z. I. A. Hakim, S. A. Fattah, Artifact: A large-scale dataset with artificial and factual images for generalizable and robust synthetic image detection, in: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, pp. 2200–2204.
[43] W. Xia, Y. Yang, J.-H. Xue, B. Wu, Tedigan: Text-guided diverse face image generation and manipulation, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 2256–2265.
[44] Z. Wang, H. Zheng, P. He, W. Chen, M. Zhou, Diffusion-gan: Training gans with diffusion (2023). arXiv:2206.02262.
[45] A. Sauer, K. Chitta, J. Müller, A. Geiger, Projected gans converge faster, in: Advances in Neural Information Processing Systems (NeurIPS), 2021.
[46] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255. doi:10.1109/CVPR.2009.5206848.
[47] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[48] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, Q. Le, Searching for mobilenetv3, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 1314–1324. doi:10.1109/ICCV.2019.00140.
[49] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 1251–1258.
[50] X. Yang, Y. Li, H. Qi, S. Lyu, Exposing gan-synthesized faces using landmark locations, in: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, Association for Computing Machinery, New York, NY, USA, 2019, p. 113–118. doi:10.1145/3335203.3335724.
URL https://doi.org/10.1145/3335203.3335724
[51] J. Wang, B. Tondi, M. Barni, An eyes-based siamese neural network for the detection of gan-generated face images, Frontiers in Signal Processing 2 (2022). doi:10.3389/frsip.2022.918725.
URL https://www.frontiersin.org/articles/10.3389/frsip.2022.918725
[52] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, International Journal of Computer Vision 128 (2) (2019) 336–359. doi:10.1007/s11263-019-01228-7.
URL http://dx.doi.org/10.1007/s11263-019-01228-7
[53] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).