Towards an Automatic System for Generating Synthetic and Representative Facial Data for Anonymization

Natália F. de C. Meira[a], Ricardo C. C. de M. Santos, Mateus C. Silva[b], Eduardo J. da S. Luz and Ricardo A. R. Oliveira[c]

Department of Computer Science, Federal University of Ouro Preto, Ouro Preto, Brazil

Keywords: GANs, Privacy, Data Anonymization, Representative Data, Face Swap, Deep Learning.

Abstract: Deep learning models based on autoencoders and generative adversarial networks (GANs) have enabled increasingly realistic face-swapping tasks. Surveillance cameras for detecting people and faces to monitor human behavior are becoming more common. Training AI models for these detection and monitoring tasks require large sets of facial data that represent ethnic, gender, and age diversity. In this work, we propose the use of generative facial manipulation techniques to build a new representative data augmentation set to be used in deep learning training for tasks involving the face. In the presented step, we implemented one of the most famous facial switching architectures to demonstrate an application for anonymizing personal data and generating synthetic data with images of drivers' faces during their work activity. Our case study generated synthetic facial data from a driver at work. The results were convincing in facial replacement and preservation of the driver's expression.

1 INTRODUCTION

Surveillance cameras that utilize pre-trained AI models are common in industrial, commercial and residential environments (Morishita et al., 2021; Eldrandaly et al., 2019). These cameras are used for various applications such as fatigue detection (Sikander and Anwar, 2018), personal protective equipment detection (Nath et al., 2020) and, human behavior classification to identify inappropriate or dangerous behavior for the individual (Ahmed and Echi, 2021; Liu et al., 2021; Alajrami et al., 2019).

Facial exchange technology has evolved with a variety of specialized approaches and techniques. Currently, the most common approaches are: face swap, and synthesized aging and rejuvenation while maintaining the person's identity. The work by Yu et al. (Yu et al., 2021) divides facial manipulation algorithms into two types: face swapping and face reenactment.

The models behind these face swap tasks are based on Autoencoders (AEs) and Adversarial Generative Neural Networks (GANs). Generative models can potentially learn any data distribution in an unsupervised way (Pavan Kumar and Jayagopal, 2021). The increasing advancement of these techniques has raised questions about privacy and manipulation of personal data for fraud, scams, and unethical and malicious applications. Mirsky et al. (Mirsky and Lee, 2021) reinforce the need to advance the detection of false content to obtain countermeasures for this content.

In this work, we propose the use of generative facial manipulation techniques for constructing a new set of data augmentation. Facial manipulation techniques for entertainment and advertising attract attention and become increasingly popular in online communities (Yu et al., 2021). With our automatic method, the user will be able to generate new representative deepfake faces by following the automatic steps shown in Figure 1.

Based on the discussion presented, we intend to develop an automatic method for generating synthetic data of representative faces for use with embedded cameras in industrial scenarios. The use of open models is a requirement. Therefore, the main objective of this text is:

- To present the method for generating representative synthetic datasets for work activities based on face-switching by Adversary Generative Neu-

[a] https://orcid.org/0000-0002-7331-6263
[b] https://orcid.org/0000-0003-3717-1906
[c] https://orcid.org/0000-0001-5167-1523

Meira, N., Santos, R., Silva, M., Luz, E. and Oliveira, R.
Towards an Automatic System for Generating Synthetic and Representative Facial Data for Anonymization.
DOI: 10.5220/0011796800003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 5: VISAPP*, pages 854-861
ISBN: 978-989-758-634-7; ISSN: 2184-4321

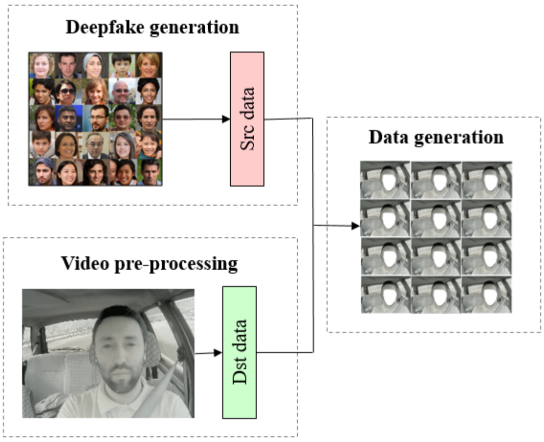

Figure 1: Representation of the automatic system for generating representative synthetic datasets.

ral Networks.

We propose to develop this work through the following steps:

- **Step 1** – Generate representative deepfake faces with ethnic, racial, gender, and age diversity to avoid biases in facial datasets;
- **Step 2** – Provide one or more videos of the situation studied for the model;
- **Step 3** – Train the model to get replicas of the videos provided with the faces swapped. These videos will provide a representative, anonymized synthetic facial dataset for training models for tasks involving facial datasets.

This work aims to explore a case study of monitoring the work activity of drivers of a transport company, with embedded cameras onboard the vehicle, to obtain a proof of concept for Step 2 and Step 3. For this, the case study consists of replacing the face of the individual during the performance of his work through the deepfake technique using GAN.

The main contribution of this work is to present a proof of concept as an alternative for generating synthetic and anonymized data representative of employees during the execution of the work activity.

This work is structured as follows: In Section 1, we presented the problem and the main contributions of the work. In Section 2 we provide a theoretical review of deepfake algorithms, metrics for evaluating GANs, and challenges for detecting and creating fake content. In Section 3 we present some works related to people detection and ethics in AI. In Section 4 we present the methodology implemented, with the results achieved in Section 5 . Finally, the conclusions are in Section 6.

2 THEORETICAL BACKGROUND

In this section, we present some concepts and research that have advanced in the context of deepfakes: evolving approaches, face swap approaches, metrics for evaluating deepfakes, and recent discussions on false content detection and future challenges.

2.1 Popularization of Face Swap Frameworks: DeepFaceLab

Most deepfakes are created using variations or combinations of generative adversarial networks and encoder-decoder networks (Mirsky and Lee, 2021; Yu et al., 2021). Among these, the GANs introduced by Goodfellow et al. (Goodfellow et al., 2014) have gained particular attention for quality imaging and data augmentation tasks (Pavan Kumar and Jayagopal, 2021).

The popularization of facial manipulation algorithms gained prominence in 2017, with the publication of a deepfake video posted by a Reddit user and, in the same year, videos of former President Barak Obama (Suwajanakorn et al., 2017; Kumar et al., 2017; Jalalifar et al., 2018).

Then, deepfake algorithms for facial swap and synthesis became popular, such as Fast Face-Swap (Korshunova et al., 2017), Faceswap-GAN (Shaoanlu, 2018), and DeepFaceLab (Perov et al., 2020).

DeepFaceLab (DFL)[1] (Perov et al., 2020) is an open-source deepfake project used to create facial manipulation videos. The three phases of the DFL pipeline are extraction, training, and conversion:

- **Extraction** – aims to extract a face from the source (src) and destination (dst) sets. It performs face detection, face alignment, and face segmentation;
- **Training** – training has two structures;
- **Conversion** – swap face from src to dst and vice versa.

At the time of its release, DFL had competitive results with other face swap frameworks in experiments under the same conditions. Currently, DFL has a support community for users' implementations and questions.

2.2 Metrics for GANs

The work of Kumar and Jayagopal (Pavan Kumar and Jayagopal, 2021) organized the main evaluation met-

[1] https://github.com/iperov/DeepFaceLab

rics for GANs. The authors found that no metrics were developed and standardized to evaluate GANs. In addition, evaluation proposals rely on a distance function to calculate the distance between an actual distribution and a generated distribution. Some of the metrics presented were:

- **Nearest Neighbor Classifier (1-NN)** – is a version of the classifier two sample tests (C2ST), and not an evaluation metric and aims to verify a similarity between the distribution of real data and the distribution of generated data;
- **Inception Scores (IS)** – this metric is derived from the work of Salimans et al. (Salimans et al., 2016), and is used to assess the quality and diversity of images synthesized by generative models;
- **Mode Score (MS)** - metric that overcomes the limitation faced by the IS metric and considers the previous distribution statistics to assess the quality and diversity of images (Cai et al., 2019).

Other metrics presented are: Maximum mean discrepancy (MMD) (Guo et al., 2020), Multi-scale structural similarity for image quality (Wang et al., 2004), and Wasserstein critical (Arjovsky et al., 2017).

2.3 False Content Detection

The work of Yu et al. (Yu et al., 2021) aimed to demonstrate the current status of deepfake video detection research. The authors cited the recent merger of Amazon, Facebook, and Microsoft, which teamed up to host the Deepfake Detection Challenge (DFDC) to build innovative technologies beneficial to detecting deepfake videos. The authors also cite that traditional methods are not suitable for detecting deepfakes.

The proposed methods for detecting deepfakes can be divided into at least five categories:

- **General-network-Based Methods:** detection is considered a frame-level classification task that CNNs complete;
- **Temporal-consistency-Based Methods:** several works have implemented RNNs to detect inconsistencies between adjacent frames;
- **Visual Artifacts-Based Methods:** discrepancies that appear at the image blend boundaries;
- **Camera-fingerprints-Based Methods:** different devices leave different traces on captured images. So it is possible to identify when face and background images are coming from different devices;
- **Biological-signals-Based Methods:** approaches based on blink rate and heart rate.

Other works aimed to use multitasking learning (Nguyen et al., 2019) , attention mechanisms (Dang et al., 2020) and methods based on visual artifacts (Li and Lyu, 2018) among others.

2.4 Challenges in Creating Deepfakes

The works of Mirsky and Lee (Mirsky and Lee, 2021), Yu et al. (Yu et al., 2021) and Kumam and Jayagopal (Pavan Kumar and Jayagopal, 2021) pointed out some of the challenges in creating and detecting realistic deepfakes:

- **Generalization** – deepfakes are data-driven and reflect training data in the output, especially for disparate data distributions;
- **Identity Leakage** - sometimes, the driver's identity is partially transferred to the final deepfake image. This problem is recurrent in facial image synthesis problems;
- **Occlusions** – occur when part of the source or target image is obstructed by a hand, hair, glasses, or any other item;
- **Temporal Coherence** – deepfake videos often produce more obvious artifacts with oscillations and irregular variations;
- **Trade-Off** – network complexity and detection effect;
- **Robustness** – detecting fake material is more difficult in compressed videos;
- **Trap for the Generator** – if the generator is not as good as the discriminator, then the discriminator always differentiates between real and fake data. This feature can be caused by gradient leakage from the generators;
- **Mode Collapse Problem of GANs** - the generator tries to over-optimize the discriminator at each epoch of the GAN training. If the discriminator is caught in the local minimum trap and always rejecting all instances of its inputs, then the generator continues to generate the same set of instances;
- **Convergence Problem** – sometimes, GAN training does not converge due to irregularities in the model structure, hyperparameter tuning, and training strategies.

These challenges point out how research still needs to advance in developing and detecting quality deepfake videos.

3 RELATED WORKS

In this Section, we present some works that reinforce the thesis about the importance of surveillance and inspection in productive and typical environments today. These systems are commonly trained with AI to detect people in these environments.

Luo et al. (Luo et al., 2020) addressed the limitations of surveillance in hazardous areas and proposed a real-time video surveillance system to detect people and industrial plant status in a hazardous area. The proposed system accurately recognized people and provided instant feedback on unsafe behavior.

Nguyen et al. (Nguyen et al., 2019) proposed a method to solve camera surveillance systems (CSS) problems, such as storage capacity and bandwidth consumption. This issue happens due to the high demand from various industries for smart manufacturing that needs to monitor for abnormal behavior or perceivable objects.

Boudjit & Ramzan (Boudjit and Ramzan, 2022) presented the research progress in developing applications for identifying and detecting people using drone camera-based convolutional neural networks (CNN). The authors used different people in the tests to validate the application's performance.

Gorospe et al. (Gorospe et al., 2021) compared the performance of different DL methods in a human detection case. The authors also developed an environment to be embedded in any general-purpose embedded system, aiming at an edge computing system. The authors used the Visual wake words (VWW) dataset for binary classification to predict whether there is a person in the image.

Zhao et al. (Zhao et al., 2019) designed EagleEye, an efficient face detector with high accuracy and speed to be implemented in popular embedded devices with low computing power. The authors demonstrated ablation studies, and EagleEye runs on ARM Cortex-A53 (Raspberry Pi1 3b+) based embedded device at 21FPS with VGA resolution input with better accuracy than methods with the same order of computational complexity.

Yang et al. (Yang et al., 2020) also focused on the challenge of matching facial recognition performance and speed in an embedded environment. The authors designed and implemented a method based on the MTCNN algorithms for feature extraction and on the RKNN model to quantify the acceleration of feature extraction. The result was facial recognition processing on the RK3399Pro platform, meeting real-time requirements.

Lv et al. (Lv et al., 2021) used the MTCNN and LCNN algorithms to design an embedded facial recognition system. The authors significantly improved the face detection and recognition performance of the embedded system.

These are just a few examples of applications that monitor and detect individuals. People detection systems must deal with factors such as ethnic diversity, racial diversity, age range, and gender.

4 METHODOLOGY

This section presents the methodology implemented for the dataset's configuration. We also present the architecture of the GAN used in this work (see in Figure 2).

4.1 Dataset

The DFL framework fits into the one-to-one paradigm, i.e., there are two types of data: source (src) and destination (dst). We obtained source (src) images from videos publicly available on YouTube. The videos aim to cover the broadest possible range of expressions and angles.

We compose the destination (dst) dataset with images monitoring the work activity of drivers of a transport company. While driving, the images from the driver were recorded inside the moving vehicle. The videos have a resolution of 640x480, with 28 frames/second.

4.2 Pipeline

The DFL workflow consists of three sequential steps: extraction, training, and conversion.

4.2.1 Extraction

The extraction consists of extracting the face from the src and dst data. We used the whole-face extraction mode in all training. We use the standard S3FD face detector.

Then the DFL performs the alignment. In developing the DFL, the authors (Perov et al., 2020) noted that facial landmarks were vital in maintaining stability. Then, the authors implemented two canonical types of facial landmark extraction algorithms, followed by a classical method of mapping and transforming point patterns (Perov et al., 2020).

4.2.2 Training

In the training phase, the authors (Perov et al., 2020) proposed two structures: the DF structure and the

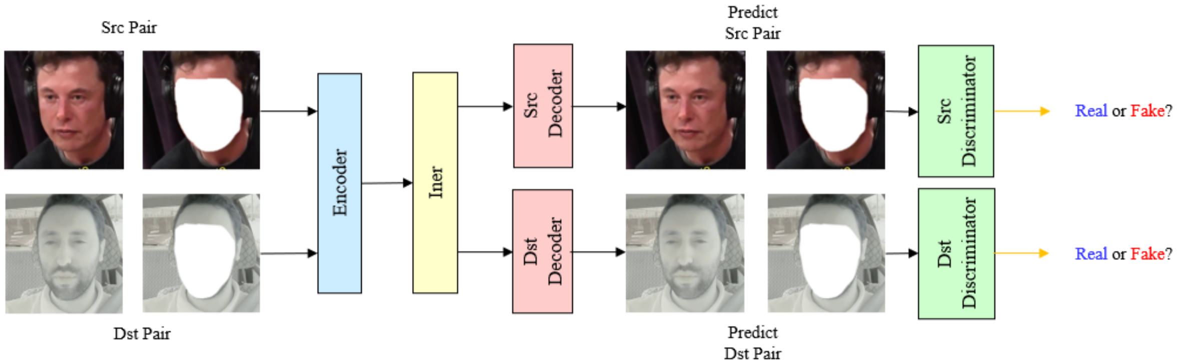

Figure 2: Overview of the extraction phase of the DF structure in DeepFaceLab (DFL) - adapted from Perov et al, (Perov et al., 2020).

LIAE structure. The architecture is composed of an Encoder and an Inter that share weights and two Decoders that belong to src and dst separately (Perov et al., 2020). The DF framework aims to solve the unpaired problem, it can complete the face swap task, but it cannot inherit enough information from dst, such as lighting.

Then, to improve the luminosity consistency problem, the LIAE structure is more complex. This structure consists of an Encoder, followed by two independent Inters. InterAB generates latent code from src and dst, while InterB generates latent code from dst only (Perov et al., 2020).

Another strategy adopted by the authors (Perov et al., 2020) is using a loss-weighted sum mask in general SSIM to improve the quality of the generated face, such as adding more weights to the eye area than the cheek area to generate a livelier look.

The standard loss consists of a mixed loss (DSSIM (structural dissimilarity) + MSE) to take advantage of both, such as: generalizing human faces faster and providing better clarity, respectively (Perov et al., 2020).

4.2.3 Conversion

The first step of converting from src to dst is transforming the generated face (see Figure 3). The face generated with the mask is reverted to the original position of the target image. Then the face is realigned to fit the target image's outer contour perfectly. Finally, a super-resolution neural network improves the sharpness and smoothing of the generated face (Perov et al., 2020).

5 RESULTS

In this Section, we present the results obtained with the training. We discuss the results related to the training data and the qualitative results in this case.

5.1 Data Results

We performed the training with the DF structure, using the Quick96 model (Perov et al., 2020), as shown in Figure 2. The training lasted about 3 hours. The hardware available for training was a computer with an AMD Ryzen 5 3600XT six-core processor, 32 GB of RAM, and the NVIDIA GeForce RTX 3090 graphics processing unit (GPU) with 21.78 GB of VRAM.

We consider different expressions and lighting conditions inside the vehicle to train the model. The image *"aligned_debug"* allows checking if the reference points are correct in the face extraction phase, as shown in Figure 3.

5.2 Qualitative Results

The model was able to perform the alignment of the front face effectively.For a lateral face, it was necessary to contemplate several faces of rotation in the data set to improve the result, but the face swap in these cases presented a realism bottleneck. We changed the available settings to achieve greater realism and alignment of the generated deepfake face in the conversion phase. This result is seen in Figure 4

Figure 5 shows some examples of the original image versus the obtained deepfake. As expected, the model had more difficulty performing face replacement when there was interference, for example, hand on face. We also observed some artifacts when the driver was fatigued, such as when the driver yawned. We hope to correct this aspect by contributing to the

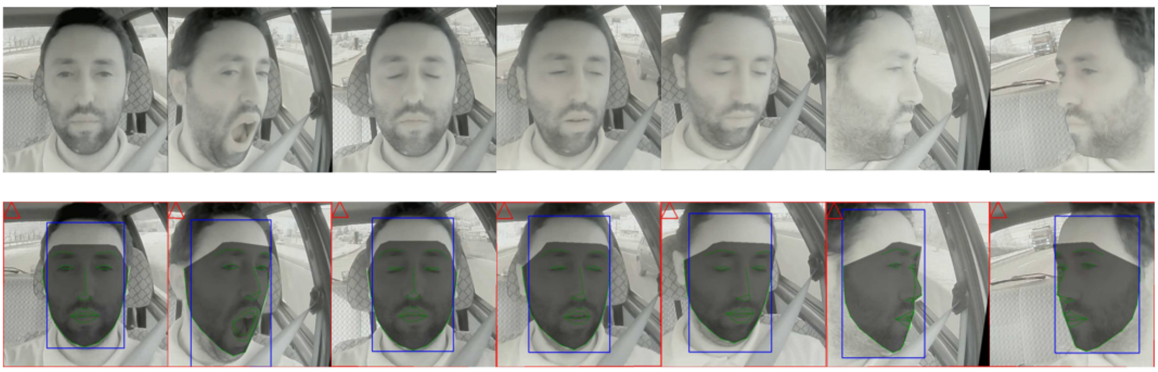

Figure 3: Images generated by ”aligned_debug” in the extraction phase for different poses and expressions. Allows the user to verify that the reference points are correct.

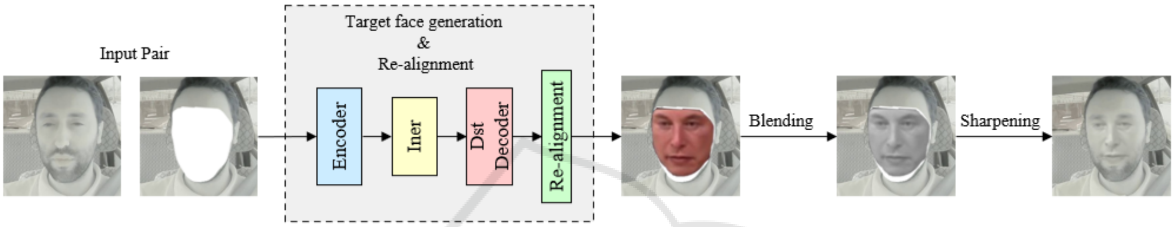

Figure 4: Overview of conversion phase in DeepFaceLab (DFL) (adapted from Perov et al., 2017).

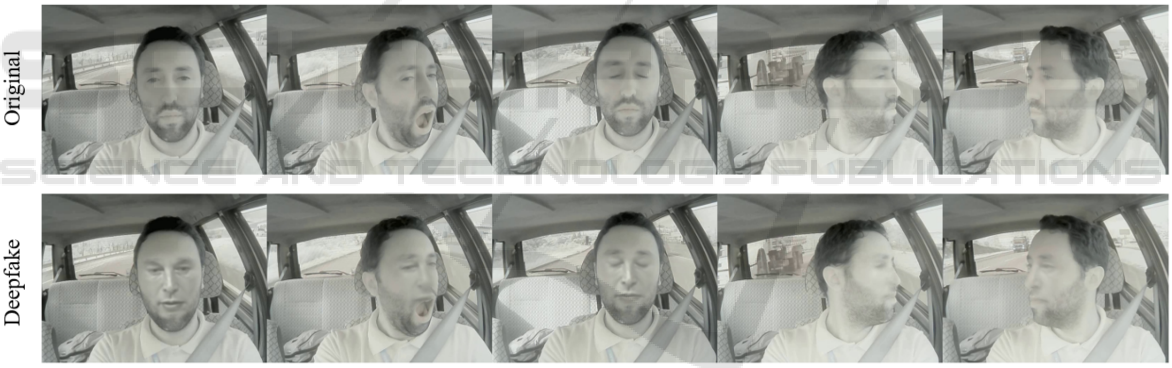

Figure 5: Original images versus deepfake synthetic images obtained with DFL training.

source dataset with more representations of the open mouth and yawning. One can see an example of the result achieved in this link: https://tinyurl.com/2o6a2g8s.

Overall, the model achieved the goal of performing the driver's anonymization while effectively preserving expressions and poses.

6 CONCLUSION

In this work, we present a proof of concept of one of the stages of development of an image data augmentation tool. The ultimate goal of this research is to generate more diverse, representative, and anonymous datasets in a GAN-based tool. In this work we present the face change phase for a work environment to generate synthetic and anonymized images.

We presented a case study an anonymization application for a vehicle driver. We built a source dataset from YouTube videos. We created a destination dataset with real vehicle driver videos. We trained the DFL model to perform face-swap and driver anonymization. We presented the quality of anonymization generated from low-resolution and low-quality images. The model provides significant anonymization results and has preserved poses and expressions.

We used the DeepFaceLab framework for this development step. We found significant difficulties with images of driver fatigue, such as yawning. In future works, we intend to improve this aspect through the dataset and test the LIAE model proposed by DFL. In the next step, we intend to develop the integration between STEP 1 and STEP 2 to automate the process of generating these datasets.

ACKNOWLEDGMENTS

The authors would like to thank CAPES, CNPq and the Federal University of Ouro Preto for supporting this work. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, the Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPQ) and the Universidade Federal de Ouro Preto (UFOP).

REFERENCES

Ahmed, A. A. and Echi, M. (2021). Hawk-eye: An ai-powered threat detector for intelligent surveillance cameras. *IEEE Access*, 9:63283–63293.

Alajrami, E., Tabash, H., Singer, Y., and El Astal, M.-T. (2019). On using ai-based human identification in improving surveillance system efficiency. In *2019 International Conference on Promising Electronic Technologies (ICPET)*, pages 91–95. IEEE.

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR.

Boudjit, K. and Ramzan, N. (2022). Human detection based on deep learning yolo-v2 for real-time uav applications. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(3):527–544.

Cai, Y., Wang, X., Yu, Z., Li, F., Xu, P., Li, Y., and Li, L. (2019). Dualattn-gan: Text to image synthesis with dual attentional generative adversarial network. *IEEE Access*, 7:183706–183716.

Dang, H., Liu, F., Stehouwer, J., Liu, X., and Jain, A. K. (2020). On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern recognition*, pages 5781–5790.

Eldrandaly, K. A., Abdel-Basset, M., and Abdel-Fatah, L. (2019). Ptz-surveillance coverage based on artificial intelligence for smart cities. *International Journal of Information Management*, 49:520–532.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, 27.

Gorospe, J., Mulero, R., Arbelaitz, O., Muguerza, J., and Antón, M. Á. (2021). A generalization performance study using deep learning networks in embedded systems. *Sensors*, 21(4):1031.

Guo, C., Huang, D., Zhang, J., Xu, J., Bai, G., and Dong, N. (2020). Early prediction for mode anomaly in generative adversarial network training: an empirical study. *Information Sciences*, 534:117–138.

Jalalifar, S. A., Hasani, H., and Aghajan, H. (2018). Speech-driven facial reenactment using conditional generative adversarial networks. *arXiv preprint arXiv:1803.07461*.

Korshunova, I., Shi, W., Dambre, J., and Theis, L. (2017). Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3677–3685.

Kumar, R., Sotelo, J., Kumar, K., de Brébisson, A., and Bengio, Y. (2017). Obamanet: Photo-realistic lip-sync from text. *arXiv preprint arXiv:1801.01442*.

Li, Y. and Lyu, S. (2018). Exposing deepfake videos by detecting face warping artifacts. *arXiv preprint arXiv:1811.00656*.

Liu, Y., Kong, L., Chen, G., Xu, F., and Wang, Z. (2021). Light-weight ai and iot collaboration for surveillance video pre-processing. *Journal of Systems Architecture*, 114:101934.

Luo, H., Liu, J., Fang, W., Love, P. E., Yu, Q., and Lu, Z. (2020). Real-time smart video surveillance to manage safety: A case study of a transport mega-project. *Advanced Engineering Informatics*, 45:101100.

Lv, X., Su, M., and Wang, Z. (2021). Application of face recognition method under deep learning algorithm in embedded systems. *Microprocessors and Microsystems*, page 104034.

Mirsky, Y. and Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)*, 54(1):1–41.

Morishita, F., Kato, N., Okubo, S., Toi, T., Hiraki, M., Otani, S., Abe, H., Shinohara, Y., and Kondo, H. (2021). A cmos image sensor and an ai accelerator for realizing edge-computing-based surveillance camera systems. In *2021 Symposium on VLSI Circuits*, pages 1–2. IEEE.

Nath, N. D., Behzadan, A. H., and Paal, S. G. (2020). Deep learning for site safety: Real-time detection of personal protective equipment. *Automation in Construction*, 112:103085.

Nguyen, H. H., Fang, F., Yamagishi, J., and Echizen, I. (2019). Multi-task learning for detecting and segmenting manipulated facial images and videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE.

Pavan Kumar, M. and Jayagopal, P. (2021). Generative adversarial networks: a survey on applications and challenges. *International Journal of Multimedia Information Retrieval*, 10(1):1–24.

Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C. S., RP, L., Jiang, J., et al. (2020). Deepfacelab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535*.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. *Advances in neural information processing systems*, 29.

Shaoanlu (2018). Faceswap-gan: A denoising autoencoder + adversarial losses and attention mechanisms for face swapping.

Sikander, G. and Anwar, S. (2018). Driver fatigue detection systems: A review. *IEEE Transactions on Intelligent Transportation Systems*, 20(6):2339–2352.

Suwajanakorn, S., Seitz, S. M., and Kemelmacher-Shlizerman, I. (2017). Synthesizing obama: learning lip sync from audio. *ACM Transactions on Graphics (ToG)*, 36(4):1–13.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612.

Yang, S., Niu, Z., Cheng, J., Feng, S., and Li, P. (2020). Face recognition speed optimization method for embedded environment. In *2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI)*, pages 147–153. IEEE.

Yu, P., Xia, Z., Fei, J., and Lu, Y. (2021). A survey on deepfake video detection. *Iet Biometrics*, 10(6):607–624.

Zhao, X., Liang, X., Zhao, C., Tang, M., and Wang, J. (2019). Real-time multi-scale face detector on embedded devices. *Sensors*, 19(9):2158.