applied sciences

MDPI

Article

Fast Fake: Easy-to-Train Face Swap Model

Tomasz Walczyna * and Zbigniew Piotrowski *

Faculty of Electronics, Military University of Technology, 00-908 Warszawa, Poland

* Correspondence: tomasz.walczyna@wat.edu.pl (T.W.); zbigniew.piotrowski@wat.edu.pl (Z.P.)

Abstract: The proliferation of “Deep fake” technologies, particularly those facilitating face-swapping in images or videos, poses significant challenges and opportunities in digital media manipulation. Despite considerable advancements, existing methodologies often struggle with maintaining visual coherence, especially in preserving background features and ensuring the realistic integration of identity traits. This study introduces a novel face replacement model that leverages a singular framework to address these issues, employing the Adaptive Attentional Denormalization mechanism from FaceShifter and integrating identity features via ArcFace and BiSeNet for enhanced attribute extraction. Key to our approach is the utilization of Fast GAN, optimizing the training efficiency of our model on relatively small datasets. We demonstrate the model’s efficacy in generating convincing face swaps with high fidelity, showcasing a significant improvement in blending identities seamlessly with the original background context. Our findings contribute to visual deepfake generation by enhancing realism and training efficiency but also highlight the potential for applications where authentic visual representation is crucial.

Keywords: face deepfake; face swap; image deepfake

1. Introduction

Deep learning has been successfully used to power applications such as big data analytics, natural language processing, signal processing, computer vision, human–computer interaction, medical imaging, and forensics. Recent advances in deep learning have also led to significant improvements in the generation of deepfakes. A deepfake consists of artificial intelligence-generated content that appears authentic to humans. The word deepfake is a combination of the words “deep learning” and “fake” and refers primarily to false content generated by an artificial neural network [1].

In the era of the growing popularity of “Deep fake” technology [2–6], more and more attention is being paid to the possibilities of content manipulation. Although fascinating, the ability to swap one face for another carries enormous potential and numerous challenges. Although many “Deep fake” techniques have already been developed and applied in various domains, from the film industry [7] to design applications, new opportunities and problems continue to emerge.

The phenomenon of deepfakes in the context of face swapping, although controversial due to potential threats to security and privacy, also opens the door to innovation and creativity in image processing. It is used in medicine [8,9], entertainment [7,10,11], and the army [12–14].

Scientists and engineers face several significant challenges in deepfake algorithms and deep learning. One key and most recognizable problem is the lack of training data. Nowadays, open training datasets are used in research. However, their scope and size are significantly limited compared to those on which current algorithms implemented in production are based. It is worth emphasizing that the size and quality of available training datasets directly and significantly impact the effectiveness and accuracy of deep learning models [15,16].

check for updates

Citation: Walczyna, T.; Piotrowski, Z. Fast Fake: Easy-to-Train Face Swap Model. *Appl. Sci.* **2024**, *14*, 2149. https://doi.org/10.3390/app14052149

Academic Editor: Douglas O’Shaughnessy

Received: 1 February 2024
Revised: 28 February 2024
Accepted: 28 February 2024
Published: 4 March 2024

Another significant challenge is the development of target-based algorithms capable of effectively distinguishing identity features from other facial attributes, such as expressions or poses [17]. This development is necessary to create reliable deepfakes that faithfully imitate the face of a reference person. Effectively separating these attributes is crucial to obtaining realistic and compelling results.

Additionally, there is the problem of matching the face to the background of the head and taking into account various occlusions. This issue requires precise algorithm work to ensure that the synthetically generated face harmoniously interacts with the surrounding environment while maintaining naturalness and visual consistency. Ensuring such a coordination between the face and the background is another challenge to effectively address when creating advanced deepfakes. Post-processing can solve this aspect, but it adds an element to the generator and slows the operation.

Another significant challenge that deep learning algorithms must face is a phenomenon known as "collapse" [18], which most often occurs during a model's training phase. This phenomenon is characterized by a situation where the generative model begins to produce a limited variety of outputs, often focusing on generating very similar or even identical instances, which significantly limits its ability to produce diverse and realistic outputs. This problem is particularly pronounced in the case of GANs, where two networks—generative and discriminative—are trained simultaneously in a competing setting. Ideally, a generative network should learn to produce data indistinguishable from the real data. In contrast, a discriminative network should learn to distinguish between real and generated data. However, in the case of a "collapse", the generative network produces a limited range of results, which leads to a loss in the diversity and quality of the generated data. This collapse, in turn, may lead to a situation in which the discriminative network quickly identifies false data, limiting the effectiveness of the entire training process [19].

In addressing critical challenges within the deepfake domain, our work introduces several key innovations that significantly advance the field. Firstly, in response to the prevalent issue of inadequate training data, our methodology circumvents the need for labelling faces under the same identity, facilitating the use of smaller datasets while maintaining a high model performance. This approach is complemented by employing a generative adversarial network (GAN) optimized for efficiency even with limited datasets. Furthermore, our novel attribute extraction method, incorporating a face segmenter within the cost function, distinctly separates identity features from facial attributes. This separation enhances the realism and fidelity of face swaps. It adeptly addresses the complex challenge of preserving background elements and occlusions, allowing for precise editing control over what aspects of an image are modified. Lastly, the implemented FastGAN [20] framework is designed to inherently resist model collapse, incorporating an additional reconstruction cost function within the discriminator for improved stability.

2. Related Works

Referring to previous work [4] on comparisons of face replacement methods, there are two main approaches to the topic: "target-based" solutions, based on extracting the identity of a person from the source image and placing it in the target image while maintaining all the characteristic features of the target image, and "source-based" approaches, in which only facial attributes in the target video/image are modified [4]. The target-based solution gives the user more control over scene development. However, its implementation is more complicated because the ground truth that can be used in training is not so obvious.

Current work on deepfake algorithms mainly revolves around three models: autoencoder, GANs, and, recently, diffusion models. Due to their "compression nature", autoencoders, in their basic version, are not the best choice for creating high-quality many-to-many deepfakes [21]. However, they are very suitable for creating one-to-one algorithms, i.e., algorithms trained for specific people [22,23]. GANs, on the other hand, are generative models that have long been considered SOTA in creating realistic samples. In GAN models, two competing models seek a balance between the discriminator and the generator. These models are susceptible to collapse, and their training is relatively slow, but they are still leaders among generative models [4,24–26]. The last model is the diffusion model, which, through the use of interconnected popular U-net models (also used in GAN models) [27], aims to generate samples through sequential denoising guided by specific conditions [28,29]. This model achieves the best results regarding generation quality and mitigating the GAN collapse problem. However, training such a model is currently very time-consuming, and the generation itself is not fast enough to be successfully used, for example, in real-time movies.

In the area of facial attribute manipulation, it is essential to distinguish between target-based and source-based methods. Target-based methods [22,24–26,29,30] allow face editing in the creator's environment, allowing users to influence the modeling process directly. In turn, source-based methods focus on adapting and changing existing materials, allowing for broad application in film post-production, education, and digital art.

Algorithms such as DeepFaceLab [22] or faceswap [23], although effective in terms of face manipulation in one-to-one models, have limitations in the context of generalization to new faces not present in the training set. Many-to-many solutions, such as FaceShifter [25] or SimSwap [24], use advanced identity-embedding techniques such as, for example, those based on the ArcFace model [31], which allows for greater flexibility in the face-swapping process. However, these methods do not verify the full range of facial attributes, such as pose and facial expressions, which may result in inauthentic or unnatural results.

An innovative approach is presented by GHOST-A [26], where a specialized cost function is introduced, focusing on maintaining the direction of gaze. This solution significantly improves realism in the generated images. However, it does not solve the problem of maintaining other key facial attributes. Moreover, integrating this function in a later phase of the training process and using additional blending and super-resolution algorithms complicate the entire process and extend the training and use times.

The DiffFace algorithm [29] presents another step forward, using diffusion techniques and relying on an additional face-parser model, ensuring the training process's stability and consistency. However, the long training time is still challenging, hindering fast production cycles.

Overall, current techniques in facial attribute manipulation still suffer from challenges related to generalization, realism, and efficiency of the training process. Although significant progress has been achieved, especially in many-to-many models and diffusion model-based techniques, many areas still require further research and development. It is essential to strive to create methods capable of quickly and reliably adjusting the entire spectrum of facial attributes, which would be a significant step forward in this dynamically developing field.

3. Methodology

As part of this research, an algorithm was developed that extends previous achievements in manipulating facial attributes. The presented method focuses on improving the direct image generation process, introducing an innovative approach to the cost function. The critical innovation of this approach consists in eliminating the need to obtain training data from the same people, significantly simplifying the data collection process and eliminating the need for detailed labelling.

Several techniques were used to achieve training acceleration, improve stability, and minimize the training data requirements of the model. These include the analysis of the hidden space in the discriminator obtained in the image reconstruction process using autoencoders [20]. Thanks to this, even with a limited amount of training data, the model can achieve satisfactory results.

The methodology's key element consists in using a pre-trained face-parser model [32,33], which enables effective image segmentation. This segmentation plays a double role: on the one hand, it supports the reconstruction of areas that should not be edited, and, on the other, it serves as an indicator for preserving key facial attributes, such as emotions or the position of the eyes, nose, and mouth. Importantly, this segmentation is not used directly in the generation process during the evaluation phase, so it does not burden the algorithm's performance.

3.1. Architecture

From the deepfake model architecture based on generative adversarial networks (GAN), we extracted two main components: the generator and the discriminator.

The central part of the architecture of our model is the generator shown in Figure 1, which consists of three key components: an attribute encoder, an identity encoder, and processing blocks. The attribute encoder uses convolutional layers with batch normalization and Leaky ReLU [34] activation functions for feature extraction. The input of the attribute encoder is a photo of the target, i.e., a photo of the face whose identity we will modify—X_t. The identity encoder uses a pre-trained face recognition model—ArcFace [31]. The input of the identity encoder in the proposed architecture is a source photo containing the face of the person whose identity we will transfer—X_s. This model provides vector representations of identity that are further transformed by fully connected layers and adapted for further processing in generating blocks. These blocks incorporate the Adaptive Attentional Denormalization (AAD) mechanism [25], which integrates pre-layers, identity embedding and attribute embedding, performing dynamic feature matching. This process is supported by residual connections [35], which enable information and gradients to propagate through the model. The generator is complemented by the use of the Skip Layer Excitation (SLE) technique [20,36], which, through channel multiplication, allows for the modulation of features generated in high-resolution blocks, contributing to a better separation of content attributes from style in the final image. The output is a ready-made photo of face Y with identity from X_s and background and attributes from X_t.

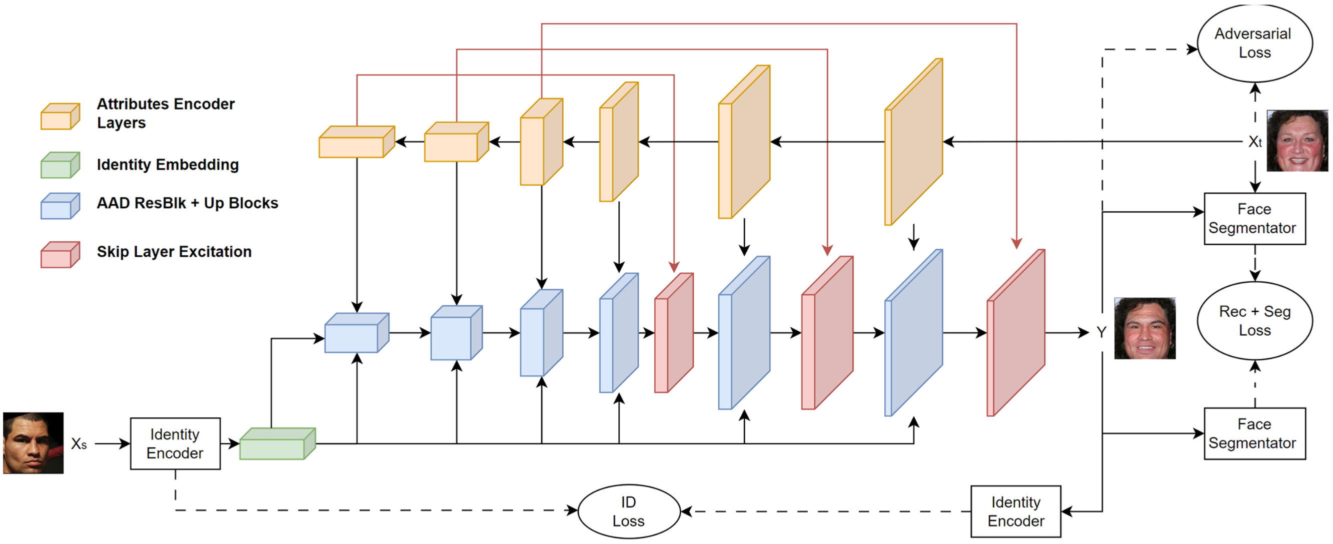

Figure 1. Generator architecture.

In the presented architecture, visible in Figure 1, attention is drawn to an additional module—the face segmenter [32]. Although it is not used directly in the production stage of image generation, it plays an invaluable role in the training phase of the model. An advanced pre-trained BiSeNet parser [27] is used for the face segmentation step. Its primary function is to identify areas that belong to the face and those that do not.

The BiSeNet parser works by analyzing the input image and classifying each pixel into the appropriate category, which allows for the accurate separation of facial structures from the background and other image elements. The information obtained from this segmentation process is crucial because it informs and improves the loss functions used during training. This implementation makes it possible to significantly improve the learning process by ensuring a more accurate match between the generated images and the actual labelled data. Moreover, it indicates the location of individual parts of the face, translating into the accuracy of maintaining the attributes. A box blur was additionally used to prevent an inaccurate mapping of structures, responsible for averaging the values of logits obtained from the parser with a specific kernel.

Later in the work, in the next chapter, specific cost functions used in the model optimization process will be discussed in more detail. It is worth noting that the appropriate definition and application of the cost function are critical for the effectiveness of learning and the quality of the final product generated by the deepfake model, a matter which emphasizes the importance of the face segmenter module in the overall architecture.

The discriminator, inspired by FastGAN [20], is designed to assess the reliability and quality of the generated images. A particular procedure for differentiable image editing [37] has been introduced into the discriminator, increasing the input data's diversity without disturbing their structural integrity. The internal structure of the discriminator contains decoders that are responsible for the process of reconstructing images labelled as authentic. Using these decoders allows for a significant increase in the effectiveness and speed of model training. By carefully refining the discrimination process, the model can subtly but precisely evaluate the generated data, which is crucial for the credibility and realism of the generated deepfakes.

3.2. Loss Functions

This chapter on the loss functions presents the critical elements in deepfake model optimization, which are intended to ensure the stability of the training process and the efficiency of image generation. Several cost functions were introduced in our model, necessary for learning complex many-to-many relationships in deepfake generation.

Identity Loss: The first one pf these functions is the so-called identity loss, which plays a fundamental role in ensuring the accuracy of the identity of the generated faces. It uses a pre-trained face recognition model, like ArcFace [31], which generates high-dimensional representations (embeddings) of faces. During training, the embeddings of the source face $\mathrm{v}_{\mathrm{X}_\mathrm{s}}$ and the generated output image v_Y are compared. Minimizing the cosine distance between these embeddings is one of the training goals, ensuring that the generated images retain vital features of the source identity.

$$\mathrm{L}_{\mathrm{Id}} = 1 - \frac{\mathrm{v}_{\mathrm{X}_\mathrm{s}} \cdot \mathrm{v}_\mathrm{Y}}{||\mathrm{v}_{\mathrm{X}_\mathrm{s}}||_2||\mathrm{v}_\mathrm{Y}||_2} \tag{1}$$

Reconstruction Loss: The second function, reconstruction loss, focuses on calculating the mean squared error (MSE) between regions that are not part of the face of the target image X_t and the generated output Y. A face segmenter such as BiSeNet is used to determine these areas, which allows for a more accurate representation of the context in which the face is located.

$$\mathrm{L}_{\mathrm{Rec}} = ||\mathrm{NotFace}_{\mathrm{mask}}(\mathrm{X}_\mathrm{t}) \times (\mathrm{X}_\mathrm{t} - \mathrm{Y})||_2 \tag{2}$$

Feature Loss: The following function, feature loss, is calculated as the mean squared error between the logits extracted by the BiSeNet segmenter network. The mean square error between the features of the X_t target image blurred with box blur and the generated output is calculated, which helps to simulate the location of facial details more accurately.

$$\mathrm{L}_{\mathrm{Feat}} = ||\mathrm{Blur}(\mathrm{Feat}(\mathrm{X}_\mathrm{t}), \mathrm{kernel}) - \mathrm{Blur}(\mathrm{Feat}(\mathrm{Y}), \mathrm{kernel})||_2 \tag{3}$$

Adversarial Loss: The essential cost function in GAN training is adversarial loss. It includes classifying images presented to the discriminator as true or false. This decision is a crucial element of the training process that pushes the generator to create more convincing images while also training the discriminator to refine their authenticity assessment.

The overall function of the generator can be described as follows:

$$\mathrm{L} = \lambda_{\mathrm{Id}}\mathrm{L}_{\mathrm{Id}} + \lambda_{\mathrm{Rec}}\mathrm{L}_{\mathrm{Rec}} + \lambda_{\mathrm{Feat}}\mathrm{L}_{\mathrm{Feat}} + \lambda_{\mathrm{Adv}}\mathrm{L}_{\mathrm{Adv}} \tag{4}$$

where $\lambda_{Id} = 5$, $\lambda_{Rec} = 100$, $\lambda_{Feat} = 40$, and $\lambda_{Adv} = 1$.

The rationale behind selecting the λs parameters in the loss function is a critical aspect of the model's design, ensuring a balanced contribution of each component to the overall training objective. The λ values guide the model to prioritize certain aspects of the image generation process, directly impacting the quality and realism of the generated deepfakes.

Identity Loss (λ_{Id}): The choice of $\lambda_{Id} = 5$ for our model was empirically determined through testing to find a sweet spot where the model sufficiently modifies the face to reflect the source identity without causing excessive alterations beyond the facial region, such as unnatural hair color changes or adding hair in inappropriate areas. This parameter's tuning is delicate; too low a value would result in minimal-to-no changes in the early training phases, whereas too high a value could introduce artefacts which detract from the realism of the face swap.

Reconstruction Loss (λ_{Rec}): The relatively high value of $\lambda_{Rec} = 100$ underscores the importance of maintaining the integrity of the background and non-facial regions of the target image. This priority ensures that, while the identity within the face region is altered, the surrounding context remains unchanged, preserving the naturalness of the overall image. The high weight of the reconstruction loss acts as a robust regularization mechanism, preventing the model from making unwanted alterations to areas outside the face.

Feature Loss (λ_{Feat}): Feature loss for our model was set to $\lambda_{Feat} = 40$, reflecting a balanced approach to integrating facial feature adjustments without overwhelming the identity and adversarial components of the model. Blurring is responsible for controlling the detail of attributes, ensuring a softer emphasis on specific facial features. This technique facilitates more natural transitions and edits, effectively simulating the subtleties of facial expressions and other transient attributes.

Adversarial Loss (λ_{Adv}): With a value of $\lambda_{Adv} = 1$, this parameter is typically kept smaller relative to the others to ensure that the generator focuses on creating convincing and authentic-looking images without overpowering the model's other objectives. The adversarial loss acts as a critical feedback mechanism, promoting the generation of images which are indistinguishable from real ones, thereby fine-tuning the model's performance in a balanced manner.

The discriminator and its classification function also use the cost function for reconstructing real images—this function is called LPIPS [38,39]. It uses latent spaces as input to decoders, which are then classified. This combined cost function ensures the stability of the training process, reducing the risk of overfitting and preventing the so-called GAN collapse, where the generator starts producing useless results.

4. Experiments

This experimental chapter will focus on the practical application and analysis of the deepfake algorithm, developed to use a pre-trained segmentation network. The basic assumption of our approach is to demonstrate this network's usefulness in separating identifying features of identity from attributes. This separation method allows for visual manipulation while maintaining key elements of personal identification.

However, this process is not without its challenges. One of the main problems we have encountered is the localization variability of a person's facial features. Faces, being dynamic and complex structures, can exhibit significant variability in different contexts and situations. Therefore, to counteract this variability, the output from the segmentation model is additionally blurred, which makes the loss function less strict. Based on a segmenter, the applied cost function shows promising results in separating and manipulating facial features in creating realistic deepfakes.

In this section, we will analyze the performance of our algorithm in detail, paying attention to its effectiveness in various scenarios and conditions. We will present a series of experiments that were performed to investigate various aspects and potential limitations of our approach. By doing so, we want to not only demonstrate the capabilities of our method but also indicate directions for future research and improvements in the field of deepfake technology.

4.1. Implementation Details

To complete the training process of our deepfake algorithm, we used the popular VGGFace2 dataset, which contains a wide range of cropped celebrity faces. This set, marked in the literature as [31], was a key data source for training and testing our model.

We used photos with a resolution of 256 × 256 pixels as the input images. We chose this size due to the optimal balance between the quality of the details and the computational requirements. Our model, inspired by FastGAN [26], is characterized by a high scalability. Thanks to this, with minor modifications, it is possible to extend our method to process images of larger dimensions, which opens the way to even more detailed and realistic results.

The features obtained from the segmenter have dimensions of 19 × 32 × 32, where 19 is the number of classes, and 32 × 32 is the size of the segmentation mask. This mask is blurred using box blur with a kernel of 15.

In the optimization context, we decided to use the Adam algorithm [40]. The parameters B1 and B2 were set to 0.9 and 0.999, respectively. This choice of parameters ensured an effective and stable learning process while minimizing the risk of getting stuck in local minima.

The batch size, i.e., the number of data samples processed in one training step, was set to 32. This choice was a compromise between training efficiency and the available computing resources.

Notably, our GAN is characterized by high training speed and stability. Thanks to this, we achieved satisfactory results after 25 training epochs, which meant carrying out about 500,000 iterations.

4.2. Results

In the results section of our study, we present the transformation effects on seven randomly selected identities that were treated in various combinations as the source or the target. An important aspect is that these specific identities were not included in the training dataset.

Figure 2 shows that our FastFake algorithm successfully transfers the identity from the source while preserving the target's attributes. This preservation is crucial to our method, allowing for realistic and reliable face manipulations.

Additionally, as part of the experiments, we compared the impact of the kernel size used on the segmentation cost function on the result. The effects of this experiment are shown in Figure 3. As the kernel size increases, more "identities" are transferred to the resulting image at the expense of attributes.

4.3. Comparison with Other Methods

Many different methods have been proposed over the years in many-to-many face replacement. We chose the most famous of them for our comparison. It should be emphasized that our goal was not to achieve the status of the SOTA (state-of-the-art) method, which would require building a larger and more complex GAN model and a much longer training process. Determining the superiority of one algorithm over another is not easy because it often depends on the specific purpose and use of the method in question. Moreover, no objective metrics would determine that a given method is better than another. As part of this article, Figure 4 compares the algorithm's results with other known ones for several selected examples.

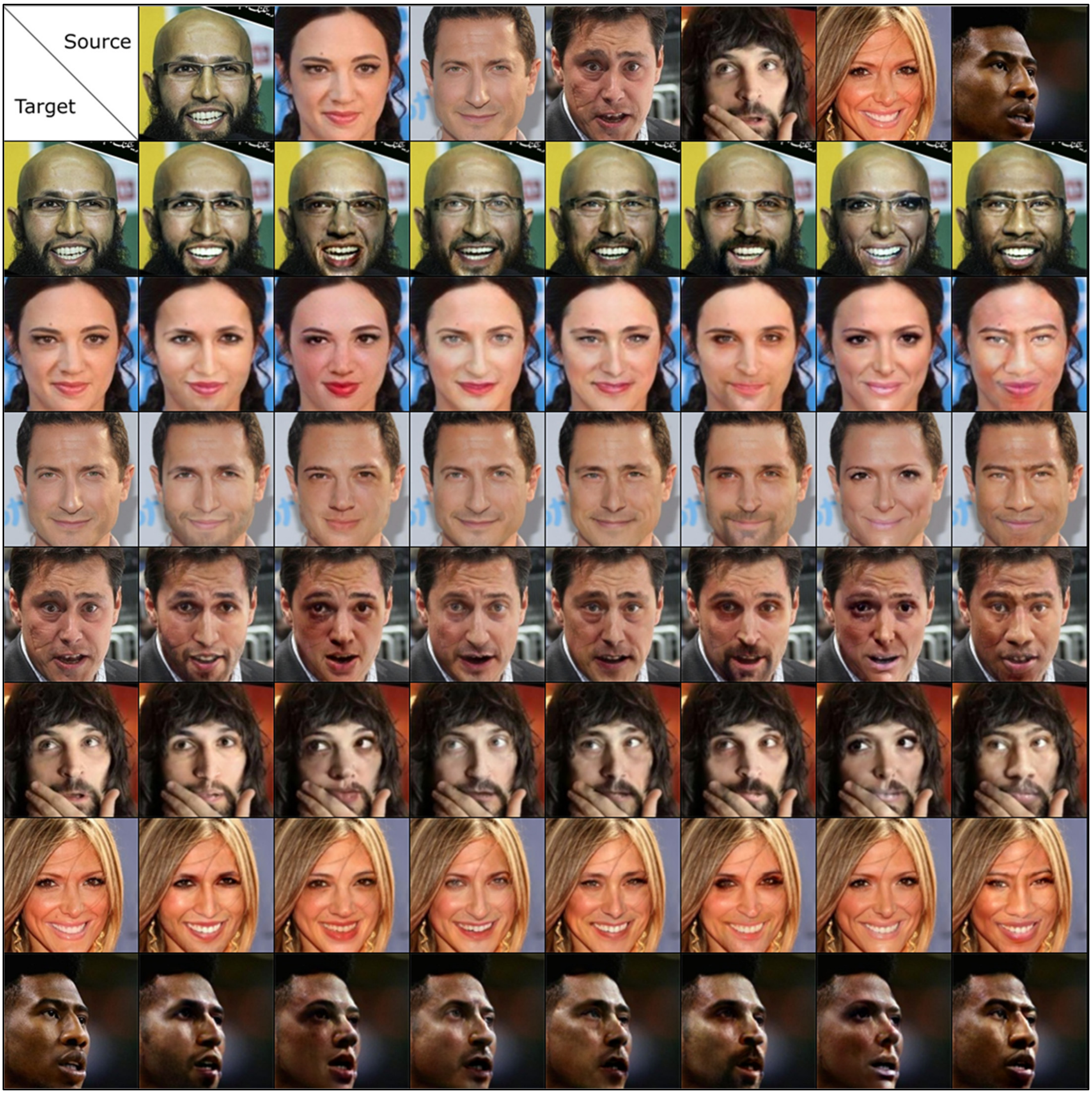

Figure 2. Results.

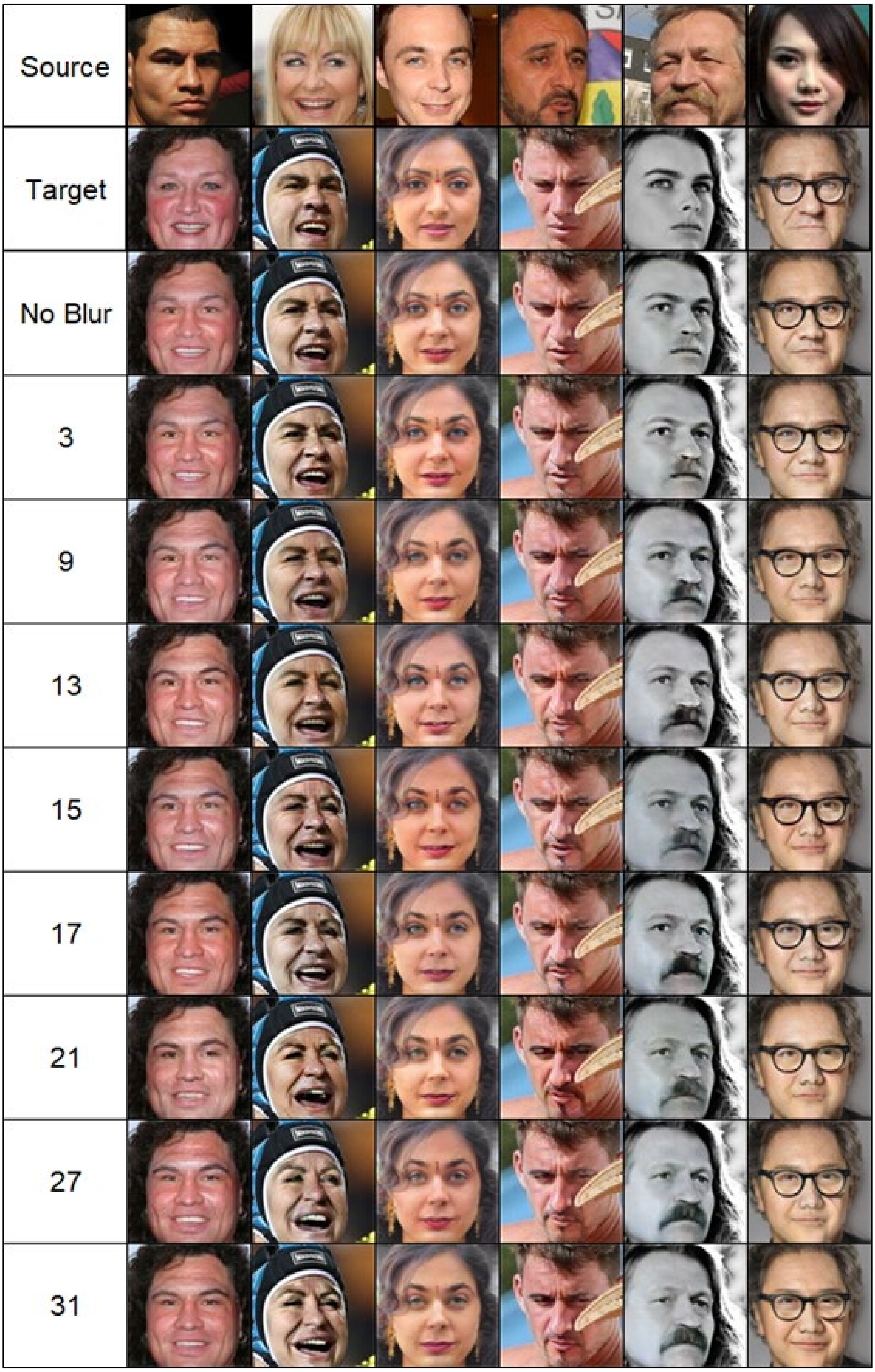

Figure 3. Results from models trained on different kernels in box blur in feature loss.

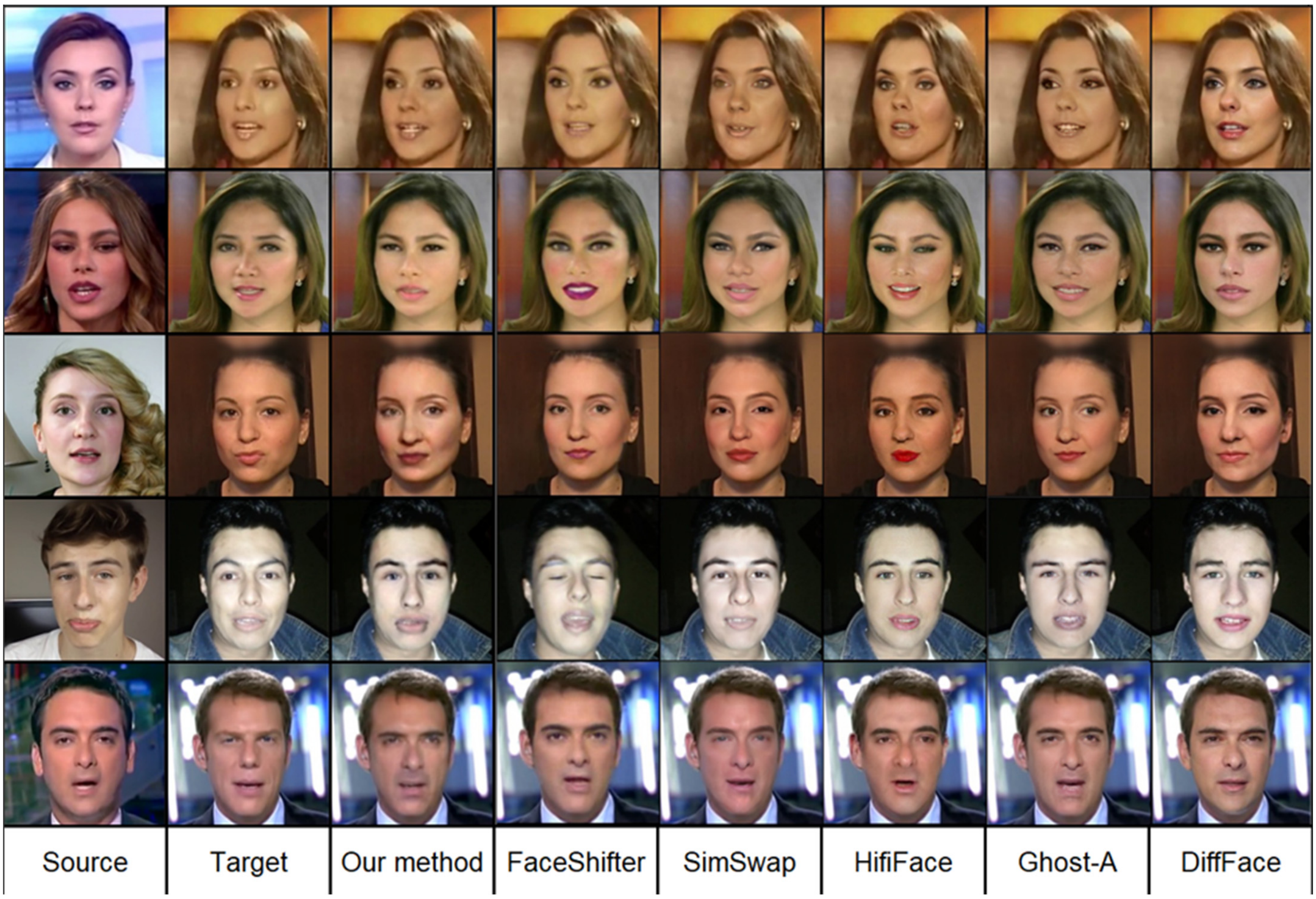

Figure 4. Comparison with other methods.

Our results are comparable to those obtained by the Ghost-A algorithm. However, our model is characterized by a shorter training time and does not require a mixed training method based on data containing the same people. Our model can be trained on data uncategorized regarding identity with comparable outcomes, which is its significant advantage.

The face-parser in our approach forces the model to edit only facial areas. It is impossible to edit an area that is not a face; this prevents additional distortions and does not require additional processing when placing a face in the target image. Using a segmenter allows for greater precision and naturalness in the face manipulation process, which is an additional advantage of our method compared to other available solutions. It can also be noted that, compared to other methods, ours is characterized by a high degree of preservation of facial attributes.

4.4. Analysis

In this section, we will focus on analyzing what our method brings to the deepfake field, its shortcomings, and what areas of development lie ahead.

Our method was primarily intended to demonstrate the effectiveness of using a pre-trained face segmentation network to preserve attributes during identity manipulation. This method showed high efficiency and realism in our experiments, preserving most potential occlusions, such as glasses, landmarks, and hair. However, we have noticed that, when training the model with a higher weight for identity embedding or using faces which are structurally significantly different, artefacts appear in the locations around the eyes and mouth; to counter this problem, we introduced a simple blur to the cost function. These artefacts can also be reduced using a different GAN model that would penalize more for less realistic generations; however, this is not a part of the current research.

The GAN used is characterized by stability and speed of training, which is a great advantage in terms of repeatability of results and ease of adaptation to new configurations. The ability to conduct many experiments with different settings is precious in research on GAN-type generative models.

Our method eliminates the need to collect data for the same people, which opens up new possibilities for expanding it with data from sets which do not segregate identities. This improvement can significantly affect the efficiency and universality of the algorithm.

There are also other areas requiring further analysis and improvement. The parser model was pre-trained on other image dimensions and only partially adapted to our needs. Considering other more advanced parsers and training them with additional elements can improve the accuracy and realism of the results. The segmentation can include other scientific achievements, such as self-attention modules [41–43], which may also improve the algorithm's operation.

Another challenge is communicating attributes such as lighting. Currently, lighting is added mainly by maintaining the realism of the result based on the background, which is always reconstructed/immutable. It is possible that extending model training to include the lighting context could improve the quality and realism of the results.

Our method significantly contributes to deepfake technology; however, many areas require further development and improvement. Future research should focus on improving the model, increasing the realism of the generated images, and better conveying subtle attributes such as lighting and gaze direction.

In the context of future works, we plan to delve into the exploration of various segmentation models that are better trained as well as investigate other models such as variational autoencoders (VAE) and diffusion models in the realm of facial subject manipulation. Additionally, we aim to transition our research focus towards video-domain applications, expanding the scope of our analysis beyond static images. This progression seeks to understand the intricacies and potential improvements in the dynamic aspects of face swapping and manipulation, leveraging the advances in segmentation and generative modeling techniques to enhance realism and applicability in video content.

5. Conclusions

In this work, we presented the FastFake method, which demonstrates the possibility of using features obtained from the segmenter in editing unstructured attributes. Our results show that this method effectively implements face replacement while preserving key attributes and occlusions such as glasses, landmarks, and hair.

The critical aspect of our method is its speed and stability of training, which allows for quick iteration and experimentation, opening the way to further innovations in this field. However, as our experiments have shown, some areas require further work and improvement, especially in the context of a more accurate transfer of attributes such as lighting and vision.

It is also worth noting that the development of deepfake technologies brings specific ethical and social challenges, especially in the context of the possibility of their abuse [41–44]. Therefore, our study highlights the importance of responsible application and further development of ethical principles in this field. Additionally, while it is not the focus of this article, it is possible to secure one's deepfake creations with unique hidden watermarks to mitigate misuse [45–47]. This is particularly important given the increasing ease of access to these technologies and their potential applications in various sectors.

In conclusion, FastFake is a promising step towards more advanced and ethically responsible deepfake technologies. Our study opens up new research and practice opportunities while highlighting the need for further research to address existing technological and ethical challenges.

Author Contributions: Investigation, T.W.; methodology, T.W.; supervision, Z.P.; validation, Z.P.; resources T.W.; writing—original draft, T.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Centre for Research and Development, with grant number CYBERSECIDENT/381319/II/NCBR/2018 on "The federal cyberspace threat detection and response system" (acronym DET-RES), as part of the second competition of the CyberSecIdent Research and Development Program—Cybersecurity and e-Identity.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Mirsky, Y.; Lee, W. The Creation and Detection of Deepfakes: A Survey. *ACM Comput. Surv.* **2022**, *54*, 1–41. [CrossRef]
2. Swathi, P.; Saritha, S.K. DeepFake Creation and Detection: A Survey. In Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), Coimbatore, India, 2–4 September 2021; pp. 584–588. [CrossRef]
3. Mahmud, B.U.; Sharmin, A. Deep Insights of Deepfake Technology: A Review. *arXiv* **2023**, arXiv:2105.00192.
4. Walczyna, T.; Piotrowski, Z. Quick Overview of Face Swap Deep Fakes. *Appl. Sci.* **2023**, *13*, 6711. [CrossRef]
5. Walczyna, T.; Piotrowski, Z. Overview of Voice Conversion Methods Based on Deep Learning. *Appl. Sci.* **2023**, *13*, 3100. [CrossRef]
6. Shahzad, H.F.; Rustam, F.; Flores, E.S.; Luís Vidal Mazón, J.; de la Torre Diez, I.; Ashraf, I. A Review of Image Processing Techniques for Deepfakes. *Sensors* **2022**, *22*, 4556. [CrossRef]
7. Usukhbayar, B. Deepfake Videos: The Future of Entertainment 2020. Master's Thesis, American University in Bulgaria, Sofia, Bulgaria, 2020. [CrossRef]
8. Yang, H.-C.; Rahmanti, A.R.; Huang, C.-W.; Li, Y.-C.J. How Can Research on Artificial Empathy Be Enhanced by Applying Deepfakes? *J. Med. Internet Res.* **2022**, *24*, e29506. [CrossRef]
9. Medical Deepfakes Are the Real Deal. Available online: https://www.mddionline.com/artificial-intelligence/medical-deepfakes-are-the-real-deal (accessed on 10 December 2023).
10. Artificial Intelligence: Deepfakes in the Entertainment Industry. Available online: https://www.wipo.int/wipo_magazine/en/2022/02/article_0003.html (accessed on 10 December 2023).
11. Caporusso, N. Deepfakes for the Good: A Beneficial Application of Contentious Artificial Intelligence Technology. In Proceedings of the AHFE 2020 Virtual Conferences on Software and Systems Engineering, and Artificial Intelligence and Social Computing, San Diego, CA, USA, 16–20 July 2020; Springer: Cham, Switzerland, 2021; pp. 235–241. [CrossRef]
12. Biddle, S.U.S. Special Forces Want to Use Deepfakes for Psy-Ops. Available online: https://theintercept.com/2023/03/06/pentagon-socom-deepfake-propaganda/ (accessed on 10 December 2023).
13. Nasu, H. Deepfake Technology in the Age of Information Warfare. Available online: https://lieber.westpoint.edu/deepfake-technology-age-information-warfare/ (accessed on 10 December 2023).
14. Bistron, M.; Piotrowski, Z. Artificial Intelligence Applications in Military Systems and Their Influence on Sense of Security of Citizens. *Electronics* **2021**, *10*, 871. [CrossRef]
15. Althnian, A.; AlSaeed, D.; Al-Baity, H.; Samha, A.; Dris, A.B.; Alzakari, N.; Abou Elwafa, A.; Kurdi, H. Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Appl. Sci.* **2021**, *11*, 796. [CrossRef]
16. Brigato, L.; Iocchi, L. A Close Look at Deep Learning with Small Data. *arXiv* **2020**, arXiv:2003.12843.
17. Olivier, N.; Baert, K.; Danieau, F.; Multon, F.; Avril, Q. FaceTuneGAN: Face Autoencoder for Convolutional Expression Transfer Using Neural Generative Adversarial Networks. *Comput. Graph.* **2023**, *110*, 69–85. [CrossRef]
18. Durall, R.; Chatzimichailidis, A.; Labus, P.; Keuper, J. Combating Mode Collapse in GAN training: An Empirical Analysis using Hessian Eigenvalues. *arXiv* **2020**, arXiv:2012.09673.
19. Thanh-Tung, H.; Tran, T. On Catastrophic Forgetting and Mode Collapse in Generative Adversarial Networks. *arXiv* **2020**, arXiv:1807.04015.
20. Liu, B.; Zhu, Y.; Song, K.; Elgammal, A. Towards Faster and Stabilized GAN Training for High-fidelity Few-shot Image Synthesis. *arXiv* **2021**, arXiv:2101.04775.
21. Zendran, M.; Rusiecki, A. Swapping Face Images with Generative Neural Networks for Deepfake Technology—Experimental Study. *Procedia Comput. Sci.* **2021**, *192*, 834–843. [CrossRef]
22. Perov, I.; Gao, D.; Chervoniy, N.; Liu, K.; Marangonda, S.; Umé, C.; Dpfks, M.; Facenheim, C.S.; RP, L.; Jiang, J.; et al. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv* **2021**, arXiv:2005.05535.
23. Deepfakes. *Deepfakes_Faceswap, v2.10.0*; GitHub: San Francisco, CA, USA, 2020. Available online: https://github.com/deepfakes/faceswap (accessed on 3 May 2023).
24. Chen, R.; Chen, X.; Ni, B.; Ge, Y. SimSwap: An Efficient Framework for High Fidelity Face Swapping. In Proceedings of the Proceedings of the 28th ACM International Conference on Multimedia, Seattle, WA, USA, 12–16 October 2020; pp. 2003–2011. [CrossRef]
25. Li, L.; Bao, J.; Yang, H.; Chen, D.; Wen, F. FaceShifter: Towards High Fidelity and Occlusion Aware Face Swapping. *arXiv* **2020**, arXiv:1912.13457.

26. Groshev, A.; Maltseva, A.; Chesakov, D.; Kuznetsov, A.; Dimitrov, D. GHOST—A New Face Swap Approach for Image and Video Domains. *IEEE Access* **2022**, *10*, 83452–83462. [CrossRef]
27. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Springer International Publishing: Cham, Switzerland, 2015; pp. 234–241. [CrossRef]
28. Dhariwal, P.; Nichol, A. Diffusion Models Beat GANs on Image Synthesis. *arXiv* **2021**, arXiv:2105.05233.
29. Kim, K.; Kim, Y.; Cho, S.; Seo, J.; Nam, J.; Lee, K.; Kim, S.; Lee, K. DiffFace: Diffusion-based Face Swapping with Facial Guidance. *arXiv* **2022**, arXiv:2212.13344.
30. Wang, Y.; Chen, X.; Zhu, J.; Chu, W.; Tai, Y.; Wang, C.; Li, J.; Wu, Y.; Huang, F.; Ji, R. HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping. *arXiv* **2021**, arXiv:2106.09965.
31. Deng, J.; Guo, J.; Yang, J.; Xue, N.; Kotsia, I.; Zafeiriou, S. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2022**, *44*, 5962–5979. [CrossRef] [PubMed]
32. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation. *arXiv* **2018**, arXiv:1808.00897.
33. Han, X.; Zhang, Z.; Ding, N.; Gu, Y.; Liu, X.; Huo, Y.; Qiu, J.; Yao, Y.; Zhang, A.; Zhang, L.; et al. Pre-Trained Models: Past, Present and Future. *arXiv* **2021**, arXiv:2106.07139. [CrossRef]
34. Xu, B.; Wang, N.; Chen, T.; Li, M. Empirical Evaluation of Rectified Activations in Convolutional Network. *arXiv* **2015**, arXiv:1505.00853.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. *arXiv* **2015**, arXiv:1512.03385.
36. Hu, J.; Shen, L.; Albanie, S.; Sun, G.; Wu, E. Squeeze-and-Excitation Networks. *arXiv* **2019**, arXiv:1709.01507.
37. Zhao, S.; Liu, Z.; Lin, J.; Zhu, J.-Y.; Han, S. Differentiable Augmentation for Data-Efficient GAN Training. *arXiv* **2020**, arXiv:2006.10738.
38. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv* **2018**, arXiv:1801.03924.
39. Lucas, A.; Tapia, S.L.; Molina, R.; Katsaggelos, A.K. Generative Adversarial Networks and Perceptual Losses for Video Super-Resolution. *IEEE Trans. Image Process.* **2019**, *28*, 3312–3327. [CrossRef]
40. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
41. Diakopoulos, N.; Johnson, D. Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections. *New Media Soc.* **2019**, *23*, 2072–2098. [CrossRef]
42. de Ruiter, A. The Distinct Wrong of Deepfakes. *Philos. Technol.* **2021**, *34*, 1311–1332. [CrossRef]
43. Karasavva, V.; Noorbhai, A. The Real Threat of Deepfake Pornography: A Review of Canadian Policy. *Cyberpsychol. Behav. Soc. Netw.* **2021**, *24*, 203–209. [CrossRef]
44. Li, M.; Wan, Y. Norms or fun? The influence of ethical concerns and perceived enjoyment on the regulation of deepfake information. *Internet Res.* **2023**, *33*, 1750–1773. [CrossRef]
45. Bistroń, M.; Piotrowski, Z. Efficient Video Watermarking Algorithm Based on Convolutional Neural Networks with Entropy-Based Information Mapper. *Entropy* **2023**, *25*, 284. [CrossRef]
46. Kaczyński, M.; Piotrowski, Z.; Pietrow, D. High-Quality Video Watermarking Based on Deep Neural Networks for Video with HEVC Compression. *Sensors* **2022**, *22*, 7552. [CrossRef]
47. Kaczyński, M.; Piotrowski, Z. High-Quality Video Watermarking Based on Deep Neural Networks and Adjustable Subsquares Properties Algorithm. *Sensors* **2022**, *22*, 5376. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.