Decoding Synthetic Face Detectors: Enhancing Interpretability with 3D Morphable Models

Giovanni Affatato, Edoardo Daniele Cannas, Sara Mandelli, Paolo Bestagini, Marco Marcon, Stefano Tubaro
Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy
giovanni.affatato@polimi.it, edoardodaniele.cannas@polimi.it, sara.mandelli@polimi.it, paolo.bestagini@polimi.it

***Abstract*—The rise of synthetic face generation and deepfakes has introduced significant challenges in multimedia forensics, particularly in ensuring the authenticity of visual content. While many detectors exist to identify these forgeries, they often operate as black boxes, providing little insight into the decision-making process. In this paper, we propose a novel method for improving the interpretability of synthetic face detectors. Our approach leverages 3D Morphable Models (3DMMs) to analyze and interpret the output of a heatmap-based detector, identifying the most important facial regions for detection. Specifically, we utilize 3DMMs to reverse-engineer the detector by averaging heatmaps from multiple images warped onto a common face geometry, revealing which areas the detector focuses on most. We validate our proposed approach by testing a state-of-the-art synthetic image detector on a dataset of real and synthetic faces generated using various Stable Diffusion (SD) models, offering insights into its behavior. Our experiments provide a valuable understanding of the internal workings of synthetic face detection, contributing to the growing need for interpretable and trustworthy forensic tools in the fight against synthetic media. Our experimental code is available at https://github.com/polimi-ispl/synthetic_image_interpretability.**

***Index Terms*—Synthetic image detection, 3D morphable models, interpretability, explainability**

I. Introduction

The rapid progress of artificial intelligence and Deep Learning (DL) has enabled the creation of highly realistic synthetic images and videos commonly known as deepfakes [1]. In particular, deep learning techniques can generate faces of unprecedented realism, threatening multimedia content's trustworthiness and integrity. The dangers we may face range from fraud to spreading fake news and revenge porn, eventually leading to loss of trust in digital content [2].

In response to the proliferation of these media, the multimedia forensics community has strongly emphasized the development of tools and techniques to detect them [1], [3], [4]. The state-of-the-art techniques are typically data-driven models that function as black boxes, providing little to no insight into the decision-making process backing the detection of synthetic content [5]. This lack of transparency hampers trust in these systems [6] and limits the ability to understand the cues that characterize manipulated content [7], [8].

This work was supported by the FOSTERER project, funded by the Italian Ministry of Education, University, and Research within the PRIN 2022 program. This work was partially supported by the European Union - Next Generation EU under the Italian National Recovery and Resilience Plan (NRRP), Mission 4, Component 2, Investment 1.3: CUP D43C22003080001, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART"); CUP D43C22003050001, partnership on "SEcurity and RIghts in the CyberSpace" (PE00000014 - program "FF4ALL-SERICS").

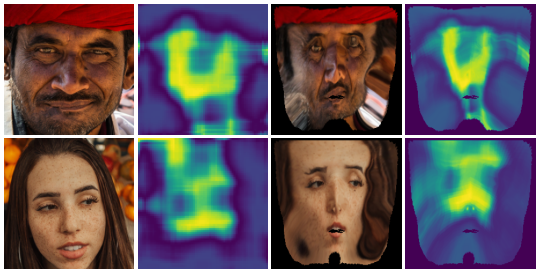


Fig. 1: Examples results of the proposed interpretability framework. Given an image (left), we return a real-valued heatmap by processing the image through a forensic detector to highlight the likelihood of every pixel being synthetically generated (second column). 3DMMs can be exploited to warp a facial image into its texture space (third column); we do such operation directly on the estimated heatmap (right), in order to align the detection scores related to facial semantics (eyes, noise, mouse, etc.) between diverse faces.

In this work, we focus on improving the interpretability of forensic detectors developed for identifying synthetically generated faces. In particular, we investigate one state-of-the-art detector and employ it to generate a 2D heatmap from the image under analysis, highlighting the likelihood of each pixel being synthetically generated. Our approach aims to provide insights into which regions of a facial image are most critical for detection. We achieve this by leveraging 3D Morphable Models (3DMMs) for representing and manipulating 3D face shapes. 3DMMs use statistical techniques to parameterize the geometrical distribution of the 3D shapes of a dataset of exemplar faces. Furthermore, this representation can be exploited to reconstruct a subject's face from a single image, making it a versatile and powerful analysis tool.

Researchers successfully employed 3DMMs in disparate applications (e.g., face recognition [9], entertainment [10], medical applications [11]), and even for generating forged content via face replacement [12] and face reenactment [13]. In multimedia forensics, the authors of [14] exploited 3DMMs to perform the decomposition of 3D reconstructed faces and fed a Convolutional Neural Network (CNN) to classify synthetic images. The authors of [15] used 3DMMs in a person-of-interest setting to classify manipulated videos with a CNN.

In the field, our work deviates from the cited articles by emphasizing the analysis of the results derived from these networks. We explore a novel way to benefit from the analytic potentiality of 3DMMs to build a forensic tool that can help shed light on the intricacies of DL methods. More

Fig. 2: Examples of 3D facial reconstruction and UV texture extraction allowed by 3DMMs. From left to right: the input facial image, its 3D facial reconstruction and the extracted UV texture.

specifically, we employ 3DMMs to reverse-engineer synthetic face detectors. By running them on facial images, we warp detection heatmaps onto a shared standard face geometry (i.e., the detection predictions related to the eyes, nose, and mouth of faces are all aligned among themselves).

Fig. 1 reports examples of our proposed interpretability framework. Given an image under analysis (on the left), we extract a real-valued detection heatmap via a selected synthetic image detector, enabling us to highlight the pixel regions with the highest detection scores (second column). By exploiting 3DMMs, we can warp each face into its texture space (third column) to move all facial key points in fixed locations. More interestingly, we can do the same operation for the extracted detection heatmaps: this operation allows us to compare detection results for multiple faces, potentially identifying which facial regions are more crucial for detection. This approach is not only widely applicable to any existing image forensic detector, but it also does not employ re-training of any parts of the pipeline on the data at hand.

To summarize, our proposed methodology aims to enhance the interpretability of face forgery detectors towards the development of more transparent forensic tools. In particular:

- we introduce a new method to analyze the results of existing synthetic face detectors via 3DMMs. This allows the inspection of facial regions for which the detector presents higher manipulation scores;
- we validate our pipeline by evaluating a state-of-the-art synthetic image detector on a dataset of synthetic faces generated through Stable Diffusion (SD) generators [3], providing some insights on its responses.

II. 3D Morphable Models

In computer vision and computer graphics, 3DMMs are generative techniques that model the 3D shape and appearance of a face [16]. The general assumption is that 3D scans of facial images are in a dense vertex-by-vertex correspondence: a vertex has the same semantic meaning across all 3D faces (e.g., the i-th vertex represents the tip of the nose on all faces). Given a dataset of exemplar 3D faces $\mathcal{S} = \{\mathbf{S}_1, \mathbf{S}_2, ..., \mathbf{S}_N\}$, each 3D shape $\mathbf{S}_i \in \mathcal{S}$ can be represented as a matrix containing in a specific order the 3D coordinates of its M vertices, i.e. $\mathbf{S}_i = [X_1^i; Y_1^i; Z_1^i, ..., X_M^i; Y_M^i; Z_M^i] \in \mathbb{R}^{3\times M}$.

The generation of a new 3D face $\mathbf{S}(\boldsymbol{\alpha}) \in \mathbb{R}^{3\times M}$ can be performed by linearly combining the example faces:

$$\mathbf{S}(\boldsymbol{\alpha}) = \mathbf{S}_1\alpha_1 + \mathbf{S}_2\alpha_2 + ... + \mathbf{S}_N\alpha_N \tag{1}$$

where $\boldsymbol{\alpha} = [\alpha_1, \alpha_2, ...\alpha_N]^T$ are the coordinates of sample faces in $\mathcal{S}$. By combining realistic examples of faces, (1) effectively models the shape space of plausible faces by intrinsically capturing the spatial distributions of each vertex's coordinates.

The geometric prior modelled by 3DMMs provides a way to perform the 3D reconstruction of a face starting from a single image. A 3D shape $\mathbf{S}(\boldsymbol{\alpha}) \in \mathbb{R}^{3\times M}$ can be projected onto the image plane by applying to each 3D vertex along its columns a known orthographic projection $\mathbf{T} \in \mathbb{R}^{2\times 3}$ and a similarity transformation, composed by a rotation $\mathbf{R} \in \mathbb{R}^{3\times 3}$, a scale factor s, and a translation $\mathbf{t}_{2\text{D}} \in \mathbb{R}^2$:

$$\mathbf{S}_{2\text{D}} = s\mathbf{TRS}(\boldsymbol{\alpha}) + \mathbf{t}_{2\text{D}}, \tag{2}$$

with $\mathbf{S}_{2\text{D}} \in \mathbb{R}^{2\times M}$. The 3D facial reconstruction problem can be formulated as estimating the parameters $\boldsymbol{\Theta} = [s, \mathbf{R}, \mathbf{t}_{2d}, \boldsymbol{\alpha}]$ given an image. Examples of 3D facial reconstructions (where the 2D points of resulting $\mathbf{S}_{2\text{D}}$ maps are shown in gray scale) are reported in the second column of Fig. 2.

3DMMs allows to perform another useful operation, which is the UV texture extraction from an image. A 2D texture is an image that stores the color information of each vertex of a 3D object. The UV mapping is the map that assign to each vertex of the 3D shape a color from 2D texture. Once estimated the parameters $\boldsymbol{\Theta}$, it is possible to sample a given facial image $\mathbf{I}$ via the unwrapping operation $U : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, which associates each 2D vertex $\mathbf{v}_{2\text{D}} = (x, y)$ along the columns of $\mathbf{S}_{2\text{D}}$ to its related texture coordinates $\mathbf{v}_{\text{UV}} = (u, v)$, such that $U(\mathbf{v}_{2\text{D}}) = \mathbf{v}_{\text{UV}}$. In specific, each pixel in the UV texture $\mathbf{I}_{\text{UV}}$ of the image $\mathbf{I}$ is defined as

$$[\mathbf{I}_{\text{UV}}]_{u,v} = [\mathbf{I}]_{x,y}. \tag{3}$$

Examples of UV textures are shown in the last column of Fig. 2. Notice that, in the UV domain of 3DMMs, each pixel represents the same semantic point across different faces. For example, the third vertex of each 3D reconstructed face is always mapped to position $(\hat{u}, \hat{v})$ for each face, independently from the actual shape. The nose tip, the eyes and the mouth are always mapped into the same UV coordinates for every input face, independently on the facial expression or subject pose. This relation sets up common ground for performing a deeper statistical analysis across all faces.

III. Proposed Methodology

Our proposed methodology is presented in Fig. 3 and is based on five main steps:

1) *Patch extraction:* given a facial image under analysis, we sequentially extract squared patches from it.
2) *Synthetic detection:* every patch is processed by a synthetic image detector to discriminate between genuine and synthetic content.

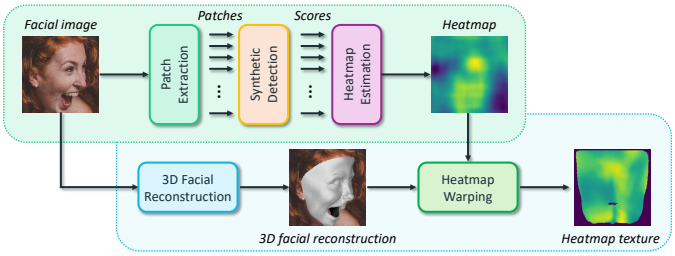


Fig. 3: Sketch of the proposed detector interpretability framework.

3) *Heatmap estimation:* patches' scores are assembled to produce a real-valued heatmap reporting the likelihood of each pixel being synthetically generated.
4) *3D facial reconstruction:* we employ 3DMMs to estimate a 3D reconstruction of the face under analysis.
5) *Heatmap warping:* we exploit 3DMMs to warp the estimated heatmap in texture coordinates.

The first three steps are independent from 3DMMs and are related to synthetic image detection. The last two stages integrate 3DMMs into the analysis process. In the following lines, we provide details for each of them.

1. Patch extraction. Given an image $\mathbf{I}$, we build our heatmaps by processing the image in patch-wise fashion, similarly to the work proposed in [17]. Specifically, given an image $\mathbf{I}$, we extract small square patches of $P \times P$ pixels sequentially. Each patch is denoted as $\mathbf{P}_{\mathbf{I}_{i,j}}$, where (i, j) represents the coordinates of the top-left corner pixel of the patch in the original image $\mathbf{I}$. Patches are extracted with a certain overlap, using a stride of $S \times S$ pixels in both dimensions.

2. Synthetic detection. Given a synthetic image detector $\mathcal{D}$, we process each patch $\mathbf{P}_{\mathbf{I}_{i,j}}$ through it, obtaining a real score $p_{i,j} = \mathcal{D}(\mathbf{P}_{\mathbf{I}_{i,j}})$ with $p_{i,j} \in [0, 1]$. This score $p_{i,j}$ represents the likelihood of the patch being synthetically generated.

3. Heatmap estimation. Patch scores are aggregated to create a real-valued heatmap that differentiates between real and synthetic pixels. We arrange the detection scores according to their corresponding patch positions in the analyzed image, resulting in an estimated tampering heatmap $\mathbf{H}$. Each score $p_{i,j}$ is linked to a patch of the heatmap $\mathbf{P}_{\mathbf{H}i,j}$, which is positioned at the same location as the corresponding patch $\mathbf{P}_{\mathbf{I}i,j}$ in the query image. The patch $\mathbf{P}_{\mathbf{H}i,j}$ is filled with constant pixel values, all equal to $p_{i,j}$. During the reconstruction process, the heatmap patches are overlaid according to the stride value. In regions where the patches overlap, the probability scores are averaged, resulting in a more refined prediction estimate.

4. 3D facial reconstruction. To perform the 3D reconstruction of the face image, we rely on 3D Dense Face Alignment Version 3 (3DDFA-V3) [18], which is a DL method based on a CNN. The architecture works in two steps. First, the input image $\mathbf{I}$ is segmented to discern characteristic parts of the face, such as mouth, eyes, nose, and so on. Then, the network regresses the parameters $\mathbf{\Theta}$ by exploiting the extracted segmentation information of the facial features.

5. Heatmap warping. We rely on the parameters $\mathbf{\Theta}$ estimated for the 3D facial reconstruction to advance the analysis of the detector's results further. For each 3D face model, instead of extracting UV textures from the image as done in the "standard" workflow reported in (3), we extract these textures directly from the heatmap $\mathbf{H}$. This process allows us to warp the extracted heatmap into a space where texture features are always aligned among faces. In other words, the warped heatmap scores related to the subject eyes will lie in the same pixel location for all the investigated faces, and similar reasoning is valid for other key points like the nose and the mouth. In doing this, (3) becomes

$$[\mathbf{H}_{\text{UV}}]_{u,v} = [\mathbf{H}]_{x,y}. \tag{4}$$

Our proposed warping operation enables us to perform a pixel-by-pixel comparison between the detection heatmaps of different faces. Indeed, each pixel in $\mathbf{H}_{\text{UV}}$ shares the same semantic meaning for all the investigated images. This feature might be very helpful in case multiple faces sharing similar characteristics (e.g., all coming from the same generator) must be addressed by a detector. In principle, a forensic analyst could exploit our proposed framework to have a clue on the general behavior of a detector over a specific dataset. For instance, the arithmetic mean between the warped heatmaps of all faces might provide useful insights on the pixel areas more prone to return high synthetic scores, thus enabling the focus of forensic investigations on a specific facial region. In our experiments, we show that the proposed strategy reveals a valid instrument for detectors' interpretability.

Furthermore, it is worth noticing that, by warping the heatmaps after the detection stage, we are preventing the introduction of interpolation artifacts that could hinder the detector's functioning and interpretability.

IV. Proposed Experimental Analysis

The proposed framework is general and can be applied with any kind of 3DMM method or forensic detector. In this section, we demonstrate its applicability with a real case scenario, i.e., the analysis of the response of a synthetic image detector on a dataset of faces. Our goal is to understand if the detector focuses on specific face attributes, e.g., eyes, mouth, ears, etc., and if this response is coherent across images generated by different techniques. Such a study is paramount to understand, for instance, if the detector is biased towards specific semantic features for a specific generator. In the following, we illustrate the setup followed in our experiments.

Dataset. The pristine face images dataset utilized for our analysis comprises 1081 subjects in different poses, with size 600×600 pixels [19]. Synthetic images are SD laundered versions of pristine data computed through SD-1.5 [20], SD-2.1 [21], SD-XL [22] and SD-XL-turbo [23] (1081 images each) by following the procedure described in [3]. More specifically, the process of SD laundering consists of passing an original image through SD encoder and decoder. The semantic content of the produced synthetic image is completely replicated, with very few details (barely visible to the human eye) potentially changed, depending on the specific autoencoder used. SD laundered images carry artifacts

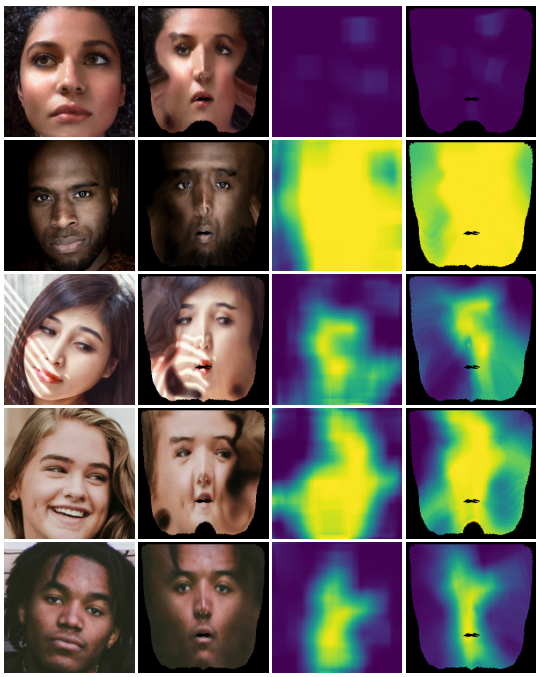


Fig. 4: Examples of the UV textures we can extract from images and heatmaps. From left to right: input image, UV textures of the input image, heatmap, UV textures of the heatmap. From top to bottom: real sample, SD-1.5, SD-2.1, SD-XL, and SD-XLTurbo.

characteristic of synthetically generated content; thus, they are typically detected as "fake" by synthetic image detectors [3], [4]. We select these synthetic images to ensure an almost perfect alignment in the semantic content between real and fake datasets. This makes our analysis independent of potential discrepancies between the semantics of the two classes.

Synthetic image detector. As a forensic detector, we select the one recently proposed in [3], which returns excellent results on the above-reported dataset (please refer to [3] for more details). Notice that while we could apply our framework to any detector potentially, i.e., even one failing in the detection task, we believe the proposed interpretation analysis is more interesting if the investigated detector achieves good classification results. Indeed, in this way, we can turn visible insights into valuable elements for forensic investigation. Given a query image, we extract 96×96 color patches with a stride of 8×8 pixels. We process each patch through the detector, returning the softmax score associated with the synthetic class. Finally, we assemble all patches' scores by following the procedure reported in Section III.

V. Results

In this section, we report the main results of our experimental analysis. Fig. 4 reports some image samples from the dataset (left column), together with the heatmaps extracted by the detector (second column) and their UV texture decomposition obtained by the 3DMM model giving the face as input (third column) and the detector's heatmap (fourth column). Every row refers to a different dataset: we start from the top with a real image and then report results for SD-1.5, SD-2.1, SD-XL and SD-XLTurbo, respectively. Blue color refers to heatmap values closer to 0, while yellow means 1.

As we can inspect, the patch-wise heatmap allows us to translate the image-level decision into a pixel-level map, indicating the areas of the faces with a higher likelihood of being synthetically generated. Given the different poses of the various subjects, a direct comparison between heatmaps is not possible. By applying the texture coordinates warping transformation instead, we can directly compare the detector's response relative to the same face attributes, e.g., mouth, eyes, forehead, etc, across different samples in various exposing conditions.

As expected, the real image shows blue values in almost all pixels, meaning for good detection accuracy in correctly classifying real pixels. The image generated through SD-1.5 shows a completely different heatmap with respect to those generated with newer SD versions. In case of SD-1.5, all facial regions seem to have the same importance for synthetic detection. This might be a sign of stronger generation artifacts, easier to detect, that have been attenuated in newer generators.

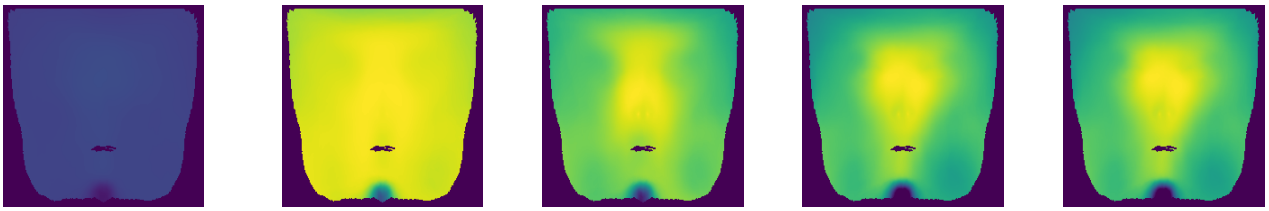
Analyzing more closely the last three rows of Fig. 4, we can gain other interesting insights. The face images of these rows have been generated through different generation techniques (i.e., different SD versions) yet the detector reports very similar responses in the same areas, namely the nose, the cheekbones and the eyebrows. This response becomes clearer thanks to the warping executed by the 3DMM, which corrects the partial face tilting of the subjects and suggests that these generators might, therefore, introduce similar artifacts that are coherently picked up by the detector in those pixel areas.

To validate our intuition, in Fig. 5 we report the UV texture heatmaps averaged across real samples and those generated by the same SD technique. These average results seem to confirm our previous considerations. In case of Fig. 5a, the detector does not focus on any specific facial attributes and does not show signs of false positives, re-assuring us of its discriminating capabilities. For all generators except SD-1.5, we can see that, on average, the central face area has the highest synthetic generation scores. Given our particular experimental setup (there are no semantic differences between the different categories of synthetic samples), the fact that the detector reports higher scores for the same facial features might be a clue for the presence of generation artifacts in that specific region. Considering that all the synthetic generation techniques in the dataset are based on the SD family, such a hypothesis, i.e., the possibility that all SD generators present systematic artifacts, is particularly fascinating. We reserve for future studies the analysis of this phenomenon.

Finally, Fig. 6 presents the results of the reverse warping of the UV texture heatmaps of Fig. 4 onto the original samples. As we can inspect, in this way, the nose areas are clearly the most detected as synthetic by the detector.

VI. Conclusions

In this paper, we presented a novel framework that enhances the interpretability of synthetic face detectors by leveraging



(a) Real. (b) SD-1.5. (c) SD-2.1. (d) SD-XL. (e) SD-XLTurbo.

Fig. 5: Average heatmaps from real and synthetically generated samples by the various generators in the dataset.

Fig. 6: Reverse warping of the UV heatmaps for Fig. 4 samples.

3DMMs. By extracting heatmaps produced by a synthetic image detector and exploiting 3DMMs for mapping them onto a common facial geometry, we were able to identify and highlight the specific facial regions that contribute most significantly to detection results. Our framework is versatile and can be applied to a wide range of detectors and 3DMM methods, making it a valuable tool for researchers and practitioners in multimedia forensics. By applying our technique, we can offer deeper insights into the patterns and biases of synthetic face detectors, ultimately contributing to the development of more transparent, interpretable, and reliable forensic tools.

In the future, we will explore the integration of our framework with more advanced 3DMMs to further enhance the precision and robustness of the interpretability process. Moreover, applying this approach to a broader set of detectors and datasets, also considering failure cases could validate its effectiveness and generalization across different synthetic media detection scenarios.

REFERENCES

[1] L. Verdoliva, "Media forensics and deepfakes: An overview," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 5, pp. 910–932, 2020.

[2] B. Paris and J. Donovan, "Deepfakes and cheap fakes," *Data & Society*, 2019.

[3] S. Mandelli, P. Bestagini, and S. Tubaro, "When synthetic traces hide real content: Analysis of stable diffusion image laundering," in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2024, pp. 1–6.

[4] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva, "Raising the bar of ai-generated image detection with clip," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 4356–4366.

[5] J. J. Bird and A. Lotfi, "Cifake: Image classification and explainable identification of ai-generated synthetic images," *IEEE Access*, vol. 12, pp. 15 642–15 650, 2024.

[6] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019. [Online]. Available: https://doi.org/10.1038/s42256-019-0048-x

[7] C. Kraetzer and M. Hildebrandt, "Explainability and interpretability for media forensic methods: Illustrated on the example of the steganalysis tool stegdetect," in *Proceedings of the 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*. SCITEPRESS - Science and Technology Publications, 2024.

[8] S. Pino, M. J. Carman, and P. Bestagini, "What's wrong with this video? comparing explainers for deepfake detection," *CoRR*, vol. abs/2105.05902, 2021. [Online]. Available: https://arxiv.org/abs/2105.05902

[9] V. Blanz, S. Romdhani, and T. Vetter, "Face identification across different poses and illuminations with a 3D morphable model," in *Proceedings of Fifth IEEE International Conference on Automatic Face Gesture Recognition*. Washington, DC, USA: IEEE, 2002, pp. 202–207.

[10] T. Weise, H. Li, L. Van Gool, and M. Pauly, "Face/Off: Live facial puppetry," in *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*. New Orleans Louisiana: ACM, Aug. 2009, pp. 7–16.

[11] A. Mueller, P. Paysan, R. Schumacher, H.-F. Zeilhofer, B.-I. Berg-Boerner, J. Maurer, T. Vetter, E. Schkommodau, P. Juergens, and K. Schwenzer-Zimmerer, "Missing facial parts computed by a morphable model and transferred directly to a polyamide laser-sintered prosthesis: An innovation study," *British Journal of Oral and Maxillofacial Surgery*, vol. 49, no. 8, pp. e67–e71, Dec. 2011.

[12] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, "Exchanging Faces in Images," *Computer Graphics Forum*, vol. 23, no. 3, pp. 669–676, Sep. 2004.

[13] J. Thies, M. Zollhöfer, M. Nießner, L. Valgaerts, M. Stamminger, and C. Theobalt, "Real-time expression transfer for facial reenactment," *ACM Transactions on Graphics*, vol. 34, no. 6, pp. 1–14, Nov. 2015.

[14] X. Zhu, H. Fei, B. Zhang, T. Zhang, X. Zhang, S. Z. Li, and Z. Lei, "Face Forgery Detection by 3D Decomposition and Composition Search," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–16, 2023.

[15] D. Cozzolino, A. Rössler, J. Thies, M. Nießner, and L. Verdoliva, "ID-Reveal: Identity-aware DeepFake Video Detection," 2020.

[16] B. Egger, W. A. P. Smith, A. Tewari, S. Wuhrer, M. Zollhoefer, T. Beeler, F. Bernard, T. Bolkart, A. Kortylewski, S. Romdhani, C. Theobalt, V. Blanz, and T. Vetter, "3D Morphable Face Models—Past, Present, and Future," *ACM Transactions on Graphics*, vol. 39, no. 5, pp. 1–38, Oct. 2020.

[17] A. Manjunath, V. Negroni, S. Mandelli, D. Moreira, and P. Bestagini, "Localization of synthetic manipulations in western blot images," *arXiv preprint arXiv:2408.13786*, 2024.

[18] Z. Wang, X. Zhu, T. Zhang, B. Wang, and Z. Lei, "3D Face Reconstruction with the Geometric Guidance of Facial Part Segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 1672–1682.

[19] C. Intelligence and Y. U. Photography (CIP) Lab, Department of Computer Science, *Real and Fake Face Detection*, 2019, https://www.kaggle.com/datasets/ciplab/real-and-fake-face-detection.

[20] Computer Vision and Learning LMU Munich, *Stable Diffusion*, 2022 (accessed June 20, 2024), https://github.com/CompVis/stable-diffusion.

[21] S. AI, *Stable Diffusion Version 2*, 2022, https://github.com/Stability-AI/stablediffusion.

[22] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, and R. Rombach, "Sdxl: Improving latent diffusion models for high-resolution image synthesis," *arXiv preprint arXiv:2307.01952*, 2023.

[23] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," *arXiv preprint arXiv:2311.17042*, 2023.