# 3D Morphable Models Meet Surface Frames for Generalizable and Robust Deepfake Detection

Giovanni Affatato[†✉], Andrea Ciamarra[⋆], Edoardo Daniele Cannas[†], Sara Mandelli[†],
Benedetta Tondi[°], Roberto Caldelli[⋆‡], Paolo Bestagini[†]
[†] *Politecnico di Milano, Milan, Italy*, [⋆] *CNIT, Florence, Italy*,
[‡] *Universitas Mercatorum, Rome, Italy*, [°] *University of Siena, Siena, Italy*
✉giovanni.affatato@polimi.it

***Abstract*—With the rapid advancements in AI-generated imagery, particularly diffusion-based models, detecting synthetic human faces has become increasingly challenging. In this paper, we introduce a synthetic face detection framework that leverages two complementary features: (i) UV textures extracted using 3D Morphable Models (3DMM) and (ii) surface frames capturing geometric structures. These modalities are fused using both feature-level and score-level fusion strategies to enhance generalization to unseen generators and robustness against post-processing operations. Experimental evaluations on diverse datasets demonstrate that our proposed method outperforms single-modality and CLIP-based approaches and provides improved generalization across different diffusion generative models, as well as improved robustness against common and strong processing operations.**

***Index Terms*—Synthetic image detection, 3D Morphable Models, Surface frame, Robust deepfake detection.**

## I. Introduction

Over the past decades, social media have enabled fast communication and sharing of multimedia contents, leading to a significant increase in their production and accessibility. This has also been possible thanks to the recent advancements in deep learning techniques. Neural Networks (NNs) can be used for a variety of tasks [1], including changing the identity of a person in an image or video by swapping its face with another subject, cloning a person's voice, or generating completely synthetic images, videos, or audio clips through simple text prompts. These last developments have been possible in great part thanks to Diffusion Models (DMs) [2], which have emerged as a revolutionary technique.

However, the availability of deep learning tools allowed malicious actors to easily spread manipulated media content, with serious consequences [1], [3] in terms of disinformation campaigns, identity theft, revenge porn, etc. There is an urgent need for tools that identify if a media is synthetically generated or, in other words, a deepfake. This is the main goal of the multimedia forensics community, which, in recent years, has proposed several deepfake image detectors [1], [4]–[6].

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").

In this paper, we specifically focus on detecting diffusion-based high-resolution images of fully generated human faces, which, from now on, we define for simplicity as deepfake images. Since DMs are available in diverse network architectures, one of the requirements for deepfake detectors is strong generalization ability, i.e., to detect a deepfake independently from the specific DM used for generating the image under analysis. Furthermore, synthetic images might also undergo several postprocessing operations. For instance, while uploaded on a social network, they might be processed with downscaling, compression, color processing, etc. All of these operations might alter the subtle forensic traces the detector needs to classify the image as a deepfake, undermining its performance.

To tackle the challenges posed by generalization and post-processing, our work introduces a synthetic face detection approach that leverages two different data modalities: (i) the facial textures extracted via 3D Morphable Models (3DMMs) [7] and (ii) the geometric structure of the image extracted via local Surface Frames (SFs) [8]. The rationale behind these features is that they provide a richer representation of the forensic content of the image, allowing at the same time to generalize better over different DM architectures and being more robust to post-processing operations than standard RGB inputs.

We thoroughly evaluate detectors that use 3DMM textures and SFs as single input modalities and propose different fusion mechanisms to exploit the information in both. Then, we compare these tools against techniques using standard RGB as input. In particular, we test our proposed solutions against state-of-the-art tools based on Contrastive Language Image Pretraining (CLIP). We run all our experiments in a cross-dataset scenario, i.e., considering generators never seen during training, and measuring the performances against common post-processing operations that images can experience in the wild. Our results show that the proposed detection methods presents better generalization and robustness to the standard RGB baseline and the current state of the art.

## II. Methodology

**Problem formulation.** In this paper we address the problem of synthetic face detection, i.e., given a query image of a face, to determine whether the image is real or a deepfake. State-of-the-art approaches tackle the problem by training NNs to provide a likelihood score $s$ indicating whether the image is

Fig. 1: Proposed input modalities. Given a face (left), 3DMMs allow to project its 3D mesh on a 2D plane (2nd column). We exploit this information, together with that of the colors, to extract the UV textures (3rd column). Last column shows the SF (z-component) extracted from the same face.

synthetic ($s > 0$) or not ($s < 0$). More formally, $s = f(\mathbf{I})$, where $f(\cdot)$ is the NN-learned function and $\mathbf{I}$ the network input, generally being a $H \times W \times 3$ RGB face image.

To increase robustness and generalization, we propose to work with different input modalities from standard RGB images, namely i) the UV texture extracted from the face image and ii) the Surface Frame (SF). Our intuition is that these modalities bring additional and complementary information to the RGB input, and we can exploit this information by fusing them to boost the performance of NN-based detectors. In the following, we provide more details on our method.

**3D Morphable Models and UV textures**. In computer graphics, 3D Morphable Models (3DMMs) are statistical models used to represent the 3D shape and appearance of faces [7]. In particular, 3DMMs enable complex tasks such as 3D facial reconstruction from a single facial image (see second column of Fig. 1). Furthermore, 3DMMs ensure that each vertex of the reconstructed 3D face retains the same semantic meaning across several different facial images (e.g., the i-th vertex represents always the tip of the nose on all faces).

With 3DMMs, it is also possible to map the surface of the 3D face onto a 2D plane to create the so-called UV texture image, defined as $\mathbf{I}_{UV}$, with the exact size of the original analyzed image $\mathbf{I}$, but where each pixel corresponds to a specific point of the 3D surface. For instance, in $\mathbf{I}_{UV}$, facial features such as the nose tip, the eyes and the mouth are mapped into the same UV coordinates for every input face, independently of the facial expression or subject pose. This alignment makes UV textures a powerful tool for conducting a detailed analysis of facial datasets. The third column of Fig. 1 provides an example of $\mathbf{I}_{UV}$.

In our work, we propose to exploit UV textures for synthetic face detection. Indeed, we believe the capability of UV textures of ensuring consistency between semantic locations across different faces might act as a sort of "regularization" term for forensic detectors. This could lead to better generalization and robustness over many post-processing operations that could alter the presence of more subtle forensic traces.

**Surface frames**. The Surface Frame (SF) [8] is a per-pixel image representation that analyses the geometry of the content in terms of surfaces and objects depicted within the scene. Formally, a SF $\mathbf{F}(i)$ at each $i$-th pixel is a $3 \times 3$ matrix of mutually orthogonal unit vectors, i.e., normals, tangents, and bitangents, defined as $\mathbf{F}(i) = [\mathbf{n}(i), \mathbf{t}(i), \mathbf{b}(i)]$, with $\mathbf{n}(i), \mathbf{t}(i), \mathbf{b}(i) \in \mathbb{R}^3$ in the x-y-z space.

A recent study [9] has shown that, while capturing an image, all the inherent scene elements, e.g. illumination, shadows, and reflections, along with the camera noises, constitute low-level details that permanently affect not only the RGB pixel values but the SFs as well. This information can be used for forensic purposes. For example, camera SFs have been used effectively to detect deepfakes in both scenarios of completely generated images [10] or synthetically inpainted pictures [9], [11].

Inspired by the work presented in [10], our paper considers SFs for the deepfake image detection task. In particular, we consider the z-component of the local SFs, defined for simplicity as $\mathbf{I}_{SF}$. We recall that, for each $i$-th image pixel, the z-component of the entire SF is a three-element vector, i.e., $[\mathbf{n}_z(i), \mathbf{t}_z(i), \mathbf{b}_z(i)]$. To obtain $\mathbf{I}_{SF}$, we concatenate these elements for each pixel, constructing $\mathbf{I}_{SF}$ as a $3 \times H \times W$ matrix, where $H$ and $W$ are the height and width of the input image, respectively. Finally, we rescale these components for each pixel in the range $[0, 255]$. This way, $\mathbf{I}_{SF}$ determines a single geometrical image representation at a pixel level. Fig. 1 provides an example of the $\mathbf{I}_{SF}$ extracted from a human face.

**Fusion strategies**. While the UV textures and SFs modalities can be used as alternatives to standard RGB inputs in NN architectures, an intriguing possibility is to fuse their information. This fusion could lead to even greater improvements in robustness to in-the-wild post-processing operations and better generalization to unknown generation techniques. To this end, we consider two approaches to fuse these features together, namely Fusion Feature (FF) and Fusion Score (FS).

Fig. 2 provides a graphical overview of the proposed fusion pipelines and the corresponding training strategies. Both pipelines begin with a modality extraction module that extracts the UV texture and the SF from an input RGB image $\mathbf{I}$. This information is then processed in parallel by identical NN backbones. The two approaches differ according to how these backbones and the features they extract are fused together to produce the final score image $s$:

- For the **FF approach** (see Fig. 2a), the two parallel streams extract deep features, which are concatenated and fed into a Multi-Layer Perceptron (MLP) that generates the final detection score. During training, weights of the MLP and the two backbone networks are jointly updated until the entire pipeline converges.
- For the **FS approach** (see Fig. 2b), training is carried out in two stages. First, the two branches are trained separately on the same dataset. Once trained, a Machine Learning (ML)-based fusion module is trained to combine the detection scores produced by the two branches, i.e., $s_{UV}$ and $s_{SF}$ for the UV textures and SFs respectively. During the training of this final module, the two NN backbones weights are frozen and no more updated.

## III. Experimental setup

**Feature extraction**. To extract UV textures and SFs, we rely on state-of-the-art solutions based on NNs. In particular, we use 3D Dense Face Alignment Version 3 (3DDFA-V3) [12] to reconstruct the 3D faces from the input images and then

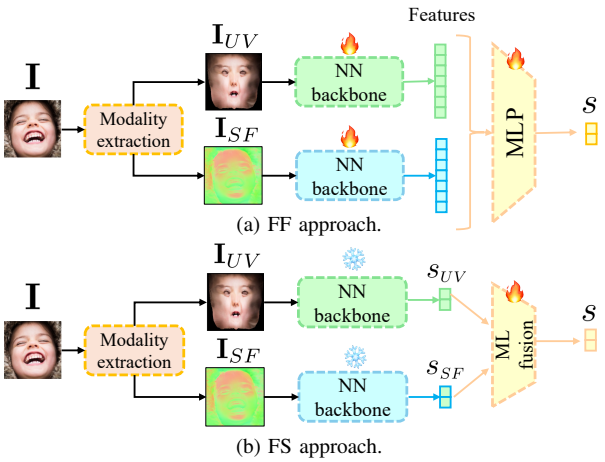


Fig. 2: Considered fusion approaches. FF trains both the UV texture and SFs backbone networks together with a MLP to predict the final deepfake score $s$. The FS trains separately the UV texture and SFs networks to predict separate scores, i.e., $s_{UV}$ and $s_{SF}$, respectively. Then, after freezing the two backbones, it trains a ML-based module to fuse the two into a final score $s$.

extract the UV textures. For SF extraction, we employ an encoder-decoder model named UpRightNet as described in following [8]. In both cases, the input image $\mathbf{I}$ must be resized to the resolution required by the respective feature extractors: $224 \times 224$ for 3DDFA-V3 and $288 \times 384$ for UpRightNet.

**Dataset**. The training set is composed of the FFHQ [13] dataset for the real faces and the recently released SFHQ-T2I [14] dataset for the fake faces. FFHQ is a well-known dataset of pristine human faces with size $1024 \times 1024$ pixels. SFHQ-T2I contains around $120K$ high-quality $1024 \times 1024$ curated face images, created through several text-to-image diffusion models, i.e., FLUX1.pro, FLUX1.dev, FLUX1.schnell, Stable Diffusion (SD)-XL, and DALL-E 3. To balance real and synthetic samples equally, we randomly select $60K$ images from both FFHQ and SFHQ-T2I for a total of $120K$ samples. Since the number of images per generator is not balanced in SFHQ-T2I, we select the samples from the two generators with more representation: $30K$ images produced by FLUX1.schnell and $30K$ images generated by SD-XL. We consider a training-validation split ratio of $5:1$.

In the testing phase, following a standard procedure in the deepfake image detection task [15], we evaluate our methods in the challenging scenario of cross-dataset generalization, meaning testing against synthetic generators that were not seen during training. We randomly pick $1K$ images for each of the remaining available generators in SFHQ-T2I, i.e., DALL-E 3, FLUX1.dev, and FLUX1.pro. In addition, we also generate about $1K$ new samples $512 \times 512$ pixels wide with the following text-to-image generation techniques: SD-1.5, SD-2.1, SD-XL1, and SD-XLTurbo. Notice that, even if SD-XL generator has been seen in training, the SFHQ-T2I dataset does not specify which version of SD-XL was used to generate the images. Our test set includes both versions SD-XL1 and SD-XLTurbo, allowing us to investigate the generalization capabilities of our detectors. Moreover, we also select completely new text prompts from those used in SFHQ-T2I. Finally, as real images, we consider the samples presented in [16], consisting of around $1K$ real human faces with size $600 \times 600$ pixels. The final test set comprises about $8K$ real and fake images.

**Networks and training details**. We employ a ResNet18 [17] pretrained on ImageNet-1K as the backbone for all our analyzes. For the FF approach, we concatenate the two embedding vectors of size 128 obtained by the two parallel branches. The MLP classifier is a fully connected layer that takes as input the fused features and outputs the final detection score $s$. We train all the networks (i.e., the two ResNet18 and the MLP module) using Adam [18] for a maximum of 100 epochs with a learning rate $\lambda = 1e^{-3}$. We train with cross-entropy loss, using early stopping with patience of 10 epochs, saving the model with the lowest validation loss. For the FS approach, as explained in Section II, we first train the two modality branches independently using the setup illustrated above. After the branches converge, we train a ML-based module on the scores extracted from the training and validation set. In particular, we employ a simple perceptron, i.e., a linear classifier trained with the Stochastic Gradient Descent (SGD) algorithm. We denote this method as FS-1. In this stage, we also consider a simpler approach based on the mean of the scores of the two branches, denoting this method as FS-2.

It is important to note that we do not include any training augmentations in our pipeline. We purposely do this to evaluate the inherent robustness and generalization capabilities of the investigated features.

**Comparison methods**. To prove the effectiveness of the proposed features, we also train a ResNet18 providing as input directly the RGB images. We follow the same training strategy as the other single-modality networks, with the only exception of the learning rate. We set $\lambda = 1e^{-4}$ as we found it to converge better and achieve better accuracies.

Finally, we also consider some recent state-of-the-art baselines, namely [4] and [5]. Both approaches use a frozen Visual Transformer [19] pretrained on the CLIP task [20] as a feature extractor and then fine-tune the network's last layer to predict the likelihood scores $s$. For our experiments, we take the models' weights provided by the original authors. Following the original papers' procedure, we then fine-tune only the last layer of the networks for 100 epochs on our training dataset using the AdamW [18] optimizer with a $\lambda = 1e^{-4}$ and reducing the learning rate on the plateau of the cross entropy loss function by a factor 0.1 with a 10 epochs patience.

As done before, the baselines are trained without any training augmentations. We do this with the specific goal of comparing the robustness and generalization of our proposed solutions with respect to those proposed in the literature.

## IV. Results

In the following, we present the results of our experimental campaign. As usually done for the deepfake image detection task, we use the Area Under the Curve (AUC) of the Receiver Operating Characteristic (ROC) and the Balanced Accuracy (BA) evaluated by thresholding at 0 the logit scores (BA@0) as comparison metrics.

TABLE I: BA evaluated at threshold 0 (%) and AUC for the cross-set experiments. In bold, the best BA per row.

| | | Proposed methodologies | | | | | State of the art | |
|---|---|---|---|---|---|---|---|---|
| Generator | RGB | UV texture | Surface frame | FF | FS-1 | FS-2 | [4] | [5] |
| DALL-E 3 | 94.90/1.00 | 96.00/0.99 | 92.90/0.99 | **97.90**/1.00 | 96.60/0.97 | 97.70/1.00 | 83.30/0.98 | 92.55/0.98 |
| FLUX1.dev | 88.80/1.00 | 98.25/1.00 | 96.90/1.00 | **99.55**/1.00 | 97.00/0.97 | 97.25/1.00 | 81.55/0.93 | 94.65/0.99 |
| FLUX1.pro | 99.40/1.00 | 98.45/1.00 | 93.60/1.00 | **99.85**/1.00 | 97.00/0.97 | 98.60/1.00 | 81.90/0.93 | 94.80/1.00 |
| SD-1.5 | 75.10/0.85 | 73.65/0.90 | 81.10/0.89 | 64.75/0.90 | **83.25**/0.83 | 80.00/0.93 | 82.45/0.95 | 77.85/0.90 |
| SD-2.1 | 78.70/0.89 | 81.30/0.94 | 84.15/0.92 | 65.35/0.92 | 89.35/0.89 | 86.30/0.96 | 82.30/0.94 | **92.00**/0.98 |
| SD-XL1 | 92.89/0.99 | 96.12/0.99 | 92.75/0.99 | 95.70/1.00 | 96.50/0.97 | **97.42**/1.00 | 83.85/0.97 | 93.51/0.99 |
| SD-XLTurbo | 86.15/0.94 | 89.10/0.97 | 89.50/0.96 | 69.65/0.96 | **93.85**/0.94 | 92.90/0.98 | 80.50/0.90 | 67.45/0.82 |
| Average | 87.99/0.95 | 90.41/0.97 | 90.13/0.96 | 84.68/0.97 | **93.36**/0.93 | 92.88/0.98 | 82.26/0.94 | 87.54/0.95 |

**Cross-set results**. In Table I, we present a comparison of AUC and BA@0 across all evaluated detection methods, with results broken down by the specific generator used in the test set.

*1) Single modalities:* As a first experiment, we compare all the single-modality networks, including RGB and our proposed features, UV texture and SF. The results are shown in the first three columns of Table I, from left to right. Focusing on each individual generators dataset, SF demonstrates the strongest generalization across all generation techniques, consistently achieving a BA above $80\%$. However, all three modalities exhibit a noticeable drop in performance when analyzing samples generated by SD-1.5 and SD-2.1. Notably, these are the oldest generators in the test set, suggesting that training on more recent generators may have led to a slight degradation in generalization to older DMs. Despite this, both UV textures and SF perform better that the RGB modality on average. These findings indicate that UV textures and SF may capture more discriminative information.

*2) Modality fusion:* Examining the fourth to sixth columns of Table I, we observe that modality fusion does not always lead to improved performance compared to single modalities. Notably, the FF pipeline exhibits a significant drop in BA when analyzing samples generated by SD-1.5, SD-2.1, and SD-XL-Turbo. However, this decline is not reflected in the AUC metric. This discrepancy suggests that the score distributions in the FF pipeline may have shifted into a range where the $0$ threshold is no longer effective in distinguishing real from deepfake images. In such cases, a dedicated score calibration process may be necessary to ensure reliable detection, particularly when encountering previously unseen generators in an in-the-wild setting. In contrast, the FS pipeline appears more robust. Across all datasets, both FS-1 and FS-2 consistently achieve BA and AUC values that are comparable to or better than those of single modalities. These results indicate that, in this case, a calibration procedure may not be required. From this perspective, the perceptron-based strategy emerges as the most effective, achieving the highest average BA.

*3) State of the art comparisons:* The CLIP-based methods proposed in [4], [5] perform worse than our proposed approaches in both the single modalities and fusion strategies. Moreover, both methods exhibit a drop in accuracy while maintaining comparable AUC, suggesting that a calibration step may be necessary to enhance their generalization.

**Robustness to post-processing**. To evaluate the robustness of our proposed methods, we apply to each image of the test set various post-processing operations. We consider editing usually applied to images in in-the-wild scenarios, e.g., social media, websites, etc. These operations include:

- Color correction: adjustments to brightness, contrast, hue, and saturation;
- Downscaling: we resize images by factors $\times 0.1$, $\times 0.25$ and $\times 0.5$; we also downscale images by $\times 0.25$ and then upscale back to their original resolution. We refer to this operation as DownUpscaling;
- JPEG compression: we compress images with quality factors of $50, 60, 70, 80$;
- Print&Scan (P&S): we simulate the printing and scanning process as done in [21].

We report the average results across all generators in Table II. We also average JPEG quality factors and downscaling levels, reporting more details on these two transforms in Fig. 3.

Color correction does not significantly impact most of the proposed methods, as the drop in BA compared to unprocessed images generally remains below $2\%$. The only method that falls below a BA of $70\%$ is the approach introduced in [4].

JPEG compression also does not pose a significant challenge for the evaluated detectors. As shown in Fig. 3, all methods perform consistently across different quality factors. This stability may be attributed to all the investigated detectors (i.e., our proposed and state of the art) always process images resized to $224 \times 224$, potentially reducing the impact of JPEG artifacts. The method proposed in [4] exhibits peak performance at a quality factor of 60, followed by a gradual decline for other factors. This behavior suggests the model may have been trained on similar compressed images.

Downscaling presents a greater challenge as a post-processing operation. Among the single modalities, SF is the most affected. However, our proposed fusion strategies improve robustness. A closer analysis of individual downscaling factors (see Fig.3) shows a clear trend: detection accuracy improves as the scale factor increases. Once again, the method proposed in [4] delivers the weakest performance. DownUpscaling proves even more challenging for the detectors. However, UV textures and the proposed fusion techniques remain robust, particularly FF and FS-2.

Notably, the P&S operation does not pose a significant challenge for our proposed detectors, while CLIP-based methods struggle with it. In general, CLIP-based approaches (e.g., [4],

TABLE II: BA@0 (%) and AUC for the Various Methods under Post-Processing Operations. In bold, the best results per column.

| Method | No Processing | Brightness | Contrast | Saturation | Hue | Downscaling | DownUpscaling | JPEG | Print&Scan |
|---|---|---|---|---|---|---|---|---|---|
| RGB | 87.99/0.95 | 87.73/0.95 | 88.09/0.94 | 86.88/0.95 | 86.68/0.95 | 81.89/0.93 | 71.83/0.90 | 87.83/0.95 | 87.09/0.95 |
| UV texture | 90.41/0.97 | 88.89/0.97 | 89.38/0.96 | 90.19/0.97 | 88.72/0.96 | 89.22/0.96 | **85.56**/0.94 | 89.83/0.97 | 86.65/0.97 |
| Surface frame | 90.13/0.96 | 85.95/0.95 | 87.18/0.95 | 89.07/0.96 | 89.56/0.96 | 76.83/0.93 | 75.23/0.91 | 89.95/0.96 | 90.01/0.96 |
| FF | 84.68/0.97 | 83.84/0.95 | 84.47/0.94 | 85.20/0.96 | 84.06/0.97 | 87.61/0.95 | 84.11/0.92 | 82.88/0.97 | 80.63/0.96 |
| FS-1 | **93.36**/0.93 | **92.34**/0.92 | 91.18/0.91 | **93.19**/0.93 | **93.05**/0.93 | 85.47/0.85 | 79.37/0.79 | **93.09**/0.93 | **92.03**/0.92 |
| FS-2 | 92.88/**0.98** | 92.10/**0.98** | **91.92**/**0.98** | 92.57/**0.98** | 92.54/**0.98** | **89.44**/**0.97** | 85.44/**0.95** | 92.00/**0.98** | 90.05/**0.98** |
| [4] | 82.26/0.94 | 76.97/0.91 | 76.37/0.90 | 81.26/0.93 | 68.27/0.90 | 63.23/0.90 | 55.69/0.86 | 80.52/0.91 | 60.43/0.85 |
| [5] | 87.54/0.95 | 84.53/0.93 | 85.31/0.93 | 86.97/0.95 | 81.64/0.92 | 83.85/0.91 | 77.48/0.86 | 82.47/0.92 | 73.39/0.88 |

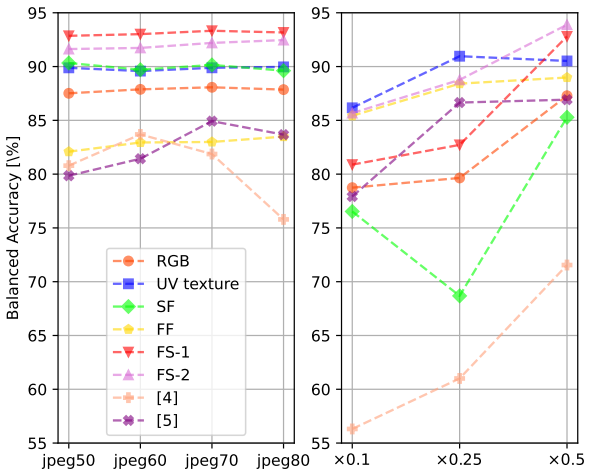


Fig. 3: BA for JPEG compressions of increasing quality factor (left) and for downscaling operations of increasing scaling factor (right).

[5]) exhibit a more substantial accuracy drop, indicating lower robustness to such operations. We believe their performance can be improved by including P&S in their training sets.

## V. Conclusions

In this work, we introduced a deepfake detection approach that exploits UV textures and surface frames to improve the generalization and the robustness performance. Our results show that leveraging the fusion of these two modalities enhances detection accuracy when the images come from unseen generators and are subject to post-processing. Among the fusion strategies tested, score fusion proved to be the most effective, outperforming traditional methods working on the RGB domain and recent techniques resorting to CLIP features.

## 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] M. Chen, S. Mei, J. Fan, and M. Wang, "An overview of diffusion models: Applications, guided generation, statistical rates and optimization," *arXiv preprint arXiv:2404.07771*, 2024.

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

[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*, 2024, pp. 4356–4366.

[5] U. Ojha, Y. Li, and Y. J. Lee, "Towards Universal Fake Image Detectors That Generalize Across Generative Models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24 480–24 489.

[6] 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.

[7] 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.

[8] W. Xian, Z. Li, M. Fisher, J. Eisenmann, E. Shechtman, and N. Snavely, "UprightNet: geometry-aware camera orientation estimation from single images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 9974–9983.

[9] A. Ciamarra, R. Caldelli, and A. Del Bimbo, "Temporal surface frame anomalies for deepfake video detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3837–3844.

[10] ——, "Spotting fully-synthetic facial images via local camera surface frames," in *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2024, pp. 1–6.

[11] ——, "Detecting deepfakes through inconsistencies in local camera surface frames," in *2024 IEEE International Conference on Image Processing Challenges and Workshops (ICIPCW)*. IEEE, 2024, pp. 4074–4080.

[12] 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.

[13] N. Research Projects, "Flickr-Faces-HQ dataset (FFHQ)," https://github.com/NVlabs/ffhq-dataset.

[14] D. Beniaguev, "Synthetic Faces High Quality - Text 2 Image (SFHQ-T2I) dataset," 2024. [Online]. Available: https://github.com/SelfishGene/SFHQ-T2I-dataset

[15] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros, "CNN-generated images are surprisingly easy to spot... for now," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8695–8704.

[16] 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.

[17] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[18] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:53592270

[19] A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, M. Minderer, M. Dehghani, N. Houlsby, S. Gelly, T. Unterthiner, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," 2021.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, 2021.

[21] N. Purnekar, L. Abady, B. Tondi, and M. Barni, "Improving the robustness of synthetic images detection by means of print and scan augmentation," in *Proceedings of the 2024 ACM Workshop on Information Hiding and Multimedia Security*, 2024, pp. 65–73.