Quantitative Metrics for Evaluating Explanations of Video DeepFake Detectors

Federico Baldassarre[1,*]
fedbal@kth.se
Quentin Debard[2]
quentin.debard@huawei.com
Gonzalo Fiz Pontiveros[2]
gonzalo.fiz.pontiveros@huawei.com
Tri Kurniawan Wijaya[2]
tri.kurniawan.wijaya@huawei.com

[1] KTH - Royal Institute of Technology
Stockholm, Sweden

[2] Huawei Ireland Research Center
Georges Court, Townsend St,
Dublin, Ireland

Abstract

The proliferation of DeepFake technology is a rising challenge in today's society, owing to more powerful and accessible generation methods. To counter this, the research community has developed detectors of ever-increasing accuracy. However, the ability to explain the decisions of such models to users is lacking behind and is considered an accessory in large-scale benchmarks, despite being a crucial requirement for the correct deployment of automated tools for content moderation. We attribute the issue to the reliance on qualitative comparisons and the lack of established metrics. We describe a simple set of metrics to evaluate the visual quality and informativeness of explanations of video DeepFake classifiers from a human-centric perspective. With these metrics, we compare common approaches to improve explanation quality and discuss their effect on both classification and explanation performance on the recent DFDC and DFD datasets.

1 Introduction

"DeepFake" refers to the realistic alteration or generation of multimedia content, in visual, audio, or textual form. The most striking application of DeepFakes are generative deep learning models that can alter a person's appearance in videos. From early attempts [75, 76], the quality of these *face-swapping* techniques has increased consistently to the point that both casual and attentive observers can be fooled. While some applications can be positively innovating [85], DeepFakes can be designed with malicious intent, such as online disinformation or public defamation. In response, the research community has introduced datasets [21, 57, 63, 65] and methods [12, 15, 36] for the automatic monitoring and detection of DeepFakes. However, benchmark performance has become the *de-facto* goal, shadowing other aspects that are crucial for the correct deployment of such models.

In practice, as automated DeepFake detectors acquire a significant role for moderation and censorship of online communities, it becomes necessary to inspect and *explain* their

[*] Work performed during an internship at Huawei Ireland Research Center

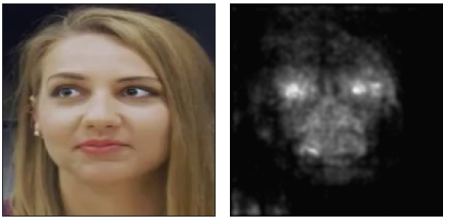


Figure 1: **Explanation heatmap** for one sample from DeepFake Dataset obtained by applying SmoothGrad to a classifier regularized with Total Variation. The resulting heatmap is visually smooth (TV $= 0.25$), localized ($\sigma = 731$), and concentrates around visible manipulation artefacts ($M_{\text{in}} = 18.1\%$), *i.e*. the eyes and one corner of the mouth.

decision process. From the users' perspective, it is not acceptable that "black-box" models manage their freedom of expression and online safety. Instead, users require intuitive explanations to validate DeepFake forgeries, prevent unjustified censorship, and trust automated moderation systems. From the perspective of companies and regulators, interpretability is necessary to justify the enforcement of DeepFake detectors, in accordance to the *right to explanation* of legal frameworks such as the GDPR [23]. Also, developers of such tools can benefit from explanations to verify the learned representation, mitigate unwanted bias, and defend against adversarial attacks.

Several methods for explaining visual classifiers exist, *e.g*. [69, 71], which can be compared in terms of faithfulness to the model and correctness to the data. However, researchers lack quantitative tools to evaluate human-centric properties of explanations and claims of improved informativeness are often based on subjective comparisons. This work introduces a quantitative framework to evaluate DeepFake explanations *w.r.t*. to human perception, which can be applied in practical deployments of DeepFake classifiers. In particular, we contextualize existing metrics, *i.e*. manipulation detection, and propose new ones as needed, namely for *smoothness*, *sparsity*, and *locality*. We apply these metrics to state-of-the-art video recognition models and compare several techniques intended for improving explanations, forming a quantitative baseline on the DeepFake Detection Challenge dataset and the DeepFake Dataset [21, 65]. Last, we empirically evaluate how to best communicate heatmap-based explanations to users, discuss limitations and future directions for DeepFake explainability.

2 Related work

DeepFake generation. Since their inception, generative models have been applied to manipulate faces, bodies and voices in online media. Today's availability of online content and ease of access to open-source frameworks, allow anyone with consumer-grade hardware to generate DeepFakes. While legitimate applications of this technology exits, *e.g*. dubbing, DeepFakes have been infamously used for disinformation, fraud, hatred, sexual abuse, and other crimes [17, 82]. This work focuses on visual forgery of faces in videos, which can be categorized as *face swapping*, in which the appearance of a face is replaced with another [19, 40, 41, 58]; or *facial reenactment*, in which expressions are edited [76, 77]. Such manipulations can be produced via purely learning-based generative models [19, 59, 77] or hybrid computer graphics approaches [40, 76]. For survey of methods and applications, we refer the reader to the works of Tolosana *et al*. [78] and Masood *et al*. [47].

DeepFake detection. In response to the widespread misuse of DeepFakes, researchers and companies have started focusing on automated detection of forged media. Forensic approaches vary from detecting anatomical inconsistencies [3, 42, 43], to analyzing digital artefacts [2, 48, 89]. Other approaches are purely learning-based [5, 30] and can integrate advanced architectures and optimization techniques [11, 15, 36, 55, 56]. Key to this effort is the release of large-scale datasets of images [18, 27, 34, 54], audio [24], and videos [21, 44, 57, 63, 65, 95], which allow to train deep models for forgery detection.

Explainable AI. Although powerful, deep learning models are often deemed "black-boxes" to illustrate the opacity of their decision process. The field of study of Explainable AI (XAI) tries to address these shortcomings to allow users, researchers, and regulators to gain insights into such models (*model interpretability*) and their outputs (*decision explainability*) [28, 60]. In the visual domain, in particular for classification, it is common to explain the decision of a model using heatmaps which highlight important areas of the input [25, 61, 91]. Backpropagation-based approaches generate heatmaps by computing gradients [63, 69, 71, 72, 74, 88, 93] or gradient surrogates [53, 68, 73]. Alternative approaches construct proxy models that are locally faithful and easier to interpret, *e.g.* LIME [64]. Recently, transformer models [7, 22, 83] have popularized using attention maps as explanations [1, 14, 87], although these might not be representative [32].

DeepFake explainability. As social platforms integrate automated tools for DeepFake detection and moderation in their pipelines [47], it becomes crucial to offer proper justification when some content is blocked. Prototype-based explanations as in Trinh *et al.* [80], could teach users to identify manipulation artefacts on their own. Similarly, SHAP-based methods can be adapted to to videos by defining 3D super-pixels [46, 62]. Focusing on input features, Wang *et al.* [84] suggest pre-processing steps that result in more human-interpretable heatmaps, according to a qualitative evaluation. Finally, human-annotated explanations, *e.g.* Mathew *et al.* [49], provide direct insight on manipulation techniques.

3 Method

3.1 Explanations methods

Our goal is to establish quantitative metrics to evaluate explanations of visual DeepFake classifiers. In particular, we focus on heatmap-based methods [8, 71, 91] that associate each pixel to a scalar proportionally to its importance *w.r.t.* the classifier decision. Formally, we define a video $v \in \mathcal{V}$ as a mapping from a discrete grid $\mathcal{G} = T \times H \times W$ to the RGB color space. A DeepFake classifier is then a function $f : \mathcal{V} \to [0,1]$ that maps a video to the probability distribution $p(\text{FAKE}|v)$. An explanation method is a function $\Phi : \mathcal{V} \times \mathcal{F} \to \mathcal{H}$ that maps a pair (v, f) to a relevance *heatmap* $h : \mathcal{G} \to \mathbb{R}^+$, where $\mathcal{F}$ and $\mathcal{H} = \{h | \int h d\lambda = 1\}$ denote the set of classifiers and heatmaps respectively. With this notation, popular gradient-based explanation methods are expressed as: Sensitivity $\nabla f(v)$ [71]; Gradient$\times$Input $\nabla f(v) \cdot v$ [82]; SmoothGrad $\mathbb{E}_{\varepsilon \sim \mathcal{N}(0,\delta I)}[\nabla f(v + v_\varepsilon)]$, where v_ε adds random color perturbations [72]; and Integrated Gradients $(v - v_b) \cdot \int_0^1 \nabla f(v_b + \alpha(v - v_b)) d\alpha$, where the baseline v_b is a uniform black video [74]. Note that ∇ and $\int$ are discretized operators over $\mathcal{G}$ (see Appendix D).

Explanation methods are commonly compared according to their *faithfulness*, *i.e.* the ability to correctly explain a decision [4, 8, 52, 64, 69]. Faithfulness is quantified by the *dele-*

tion score [61, 67] defined as $\mathbb{E}_v[\int f(v \odot (1-h_\alpha))d\alpha]$, where h_α is a binary mask obtained by selecting the most important pixels from the explanation $\Phi(f,v)$ such that their cumulative relevance is $\alpha \in [0,1]$. A low deletion score indicates a faithful explanation method: if relevant pixels are masked out first, the prediction confidence should drop sharply. For our baseline model, SmoothGrad achieves the lowest deletion score (paired one-sided t-test $p < 10^{-5}$) and is therefore selected for all evaluations of visual quality. We report per-method hyperparameters and per-dataset scores in Appendix D.4. Clearly, faithfulness is a necessary property of explanation methods, however, their heatmaps can still appear noisy and uninformative for humans, hence the need for quantitative metrics of visual quality.

3.2 Evaluation metrics

As discussed in Section 2, several works address the representativeness or visual appearance of heatmaps. However, the improvement is often demonstrated through qualitative examples, while quantitative comparison is lacking. Understandably, defining general-purpose metrics for quantifying explanation properties is not trivial [6], as the perceived quality depends on the data itself, on the target user, and the downstream task. Focussing on explanations of video DeepFake classifiers, we discuss a set of desirable human-centric properties [10, 38, 51] and formulate quantitative metrics for their evaluation.

3.2.1 Visual quality

The first set of metrics considers general properties of explanation heatmaps that facilitate their understanding and communicability. Complex models can take decisions based on features that are not easily accessible to users, *e.g.* texture details or high-frequency patterns [26, 84]. Instead, we expect models that focus on *human-interpretable cues* [38] such as small manipulation artefacts, teeth misalignment, non-circular pupils, or irregular skin complexion, to produce *smoother*, *sparser* and more *localized* heatmaps.

Smoothness. Explanations that vary excessively between neighboring pixels or frames are not meaningful to humans [84]. The smoothness of a heatmap $h : \mathcal{G} \to \mathbb{R}^+$ is measured as its Total Variation (TV), where low values indicate higher local consistency:

$$\mathrm{TV}(h) = \int_{\mathcal{G}} \|\nabla h\|_1 d\lambda. \tag{1}$$

Spatial locality. Unambiguous explanations should concentrate on few spatially-close patches of a video, *i.e.* their relevance should be localized. If we consider h as the distribution of a random vector $\boldsymbol{\rho} \in \mathcal{G}$, we can measure locality through the volume of its variance matrix:

$$\sigma = |\det(\boldsymbol{\Sigma})| = \left|\det\left(\mathbb{E}_h\left[\boldsymbol{\rho}\boldsymbol{\rho}^T\right] - \mathbb{E}_h[\boldsymbol{\rho}]\,\mathbb{E}_h\left[\boldsymbol{\rho}^T\right]\right)\right|. \tag{2}$$

A low σ will favor sharp unimodal distributions, *e.g.* a Gaussian with low dispersion, as opposed to scattered multimodal heatmaps. In the context of DeepFakes, this means highlighting single manipulation artefacts instead of allocating mass to distinct parts of the face. For other tasks, spatial locality can be extended to account for domain-specific requirements.

Sparsity. While TV and σ capture spatial properties, the individual values shall also be sparse, since few highly important regions are more indicative of a good explanation than several mildly relevant ones. Both L_0 norm and Entropy [70] are popular measures of sparsity, but the Gini Index [29] is preferred according to Hurley and Rickard [31]. For a heatmap $h : \mathcal{G} \to \mathbb{R}^+$ and sorting indices $i = \{1, \dots, THW\}$ such that $h(\boldsymbol{\rho}_i) \leq h(\boldsymbol{\rho}_{i+1})$:

$$\mathrm{G} = \frac{2}{THW} \frac{\sum_i i \cdot h(\boldsymbol{\rho}_i)}{\sum_i h(\boldsymbol{\rho}_i)} - \frac{THW+1}{THW}. \tag{3}$$

3.2.2 Manipulation detection

Smooth, sparse and localized heatmaps appear visually appealing, but do they convey the location of manipulation cues? Offering specific evidence greatly increases trust in the model, helps diagnosing failure cases, and encourages users to develop a critical eye for spotting DeepFakes. In the XAI literature, manipulation detection is commonly evaluated through user studies [59, 84], which suffers from reproducibility issues, or under a weakly-supervised paradigm [9, 13, 39, 59], which risk introducing bias from the additional annotations.

We argue that DeepFakes offer a unique possibility for the objective evaluation of weakly-supervised *manipulation detection*. Given a real video v_R, its fake(s) v_F, and a face parsing model $s : \mathcal{G} \to \mathcal{P}$ that maps pixels of v_R to $\mathcal{P} = \{\text{eyes}, \text{nose}, \text{mouth}\}$, an *ad-hoc* evaluation sample can be produced such that the manipulation is limited to a specific semantic region:

$$v_p(\boldsymbol{\rho}) = \begin{cases} v_F(\boldsymbol{\rho}) & \text{if } s(\boldsymbol{\rho}) = p \\ v_R(\boldsymbol{\rho}) & \text{otherwise} \end{cases} \qquad \forall p \in \mathcal{P} \tag{4}$$

Assuming a well-trained detector and a faithful explanation method, heatmaps for v_p should closely match the manipulated region. Since an objective ground-truth is available by construction, it's possible to assess *manipulation detection* using common segmentation metrics. First, M_{in} measures the percentage of heatmap mass inside the ground-truth mask, *i.e.* $\int_{\mathcal{G}} m_p(\boldsymbol{\rho}) h(\boldsymbol{\rho}) d\lambda$, to ensure that little or no relevance is assigned to non-manipulated regions. Second, precision at 100 (P_{100}), *i.e.* the fraction of the 100 most relevant pixels that falls inside the ground-truth, accounts for manipulation artefacts significantly smaller than the selected region. Additional manipulation detection metrics are reported in Appendix D.5.

As a point of discussion, both humans and computers may "look at" other parts of a video to assess whether one portion is manipulated, *e.g.* noting the mismatch between a smiling mouth and two frowning eyes. However, when asking "why is the video fake?", we expect to be pointed at the visible manipulation and not at other natural-looking features. Therefore, in this context, manipulation masks are considered as the ground-truth explanation.

4 Experiments

The previous section establishes a set of desirable qualities of explanations and proposes evaluation metrics built on sound mathematical foundations. We now consider several techniques from previous works and *quantify* their effect on explanations using these metrics. Section 4.1 analyses the effects of: i) data preparation [84]; ii) loss-based regularization [66]; iii) augmentation-based regularization [20]; and iv) model architecture [26, 81]. Both and classification performance (Tab. 1) and explanation quality (Fig. 2) are reported for each experiment. Furthermore, Section 4.2 discusses post-processing techniques for heatmap visualization, which are important for communicating explanations to users in practice.

Training dataset. All models are trained on videos from the DeepFake Detection Challenge [21] in "high-quality" compression (constant rate quantization 23). Specifically, we train on 19k real and 100k fake videos, and use the official validation split of 2k real and 2k fakes for hyper-parameter tuning. Each video is preprocessed using the MTCNN face detector [92], the main face is heuristically determined among all detections, then cropped and resized to 224×224 pixels. Part segmentation is obtained with the BiSeNet face parser [90] and aggregated into *background, face, nose, mouth, eyes, ears*. Additional details on data preparation and dataset statistics are provided in Appendix A.

Explanation datasets. For a cross-dataset evaluation of explanation quality metrics we employ a held-out subset of DFDC, which has a distribution similar to training videos, and a subset of the DeepFake Detection Dataset (DFD)[57], which is more challenging due to the potential distribution shift. Visual quality metrics (Sec. 3.2.1) are computed on the explanations of fake videos, while manipulation detection (Sec. 3.2.2) is evaluated on three part-swaps per video, namely *eyes*, *mouth* and *nose*. Notably, manipulation detection can only be evaluated on a subset of temporally and spatially aligned videos due to the part-swapping procedure. Additional details are provided in Appendix A. While the proposed metrics can be flexibly applied to any dataset of real-fake video pairs, we release the code for preprocessing, training and evaluation on DFDC and DFD to encourage comparison and facilitate reproducibility: github.com/baldassarreFe/deepfake-detection.

Classifier. Our baseline model is a 3D CNN trained with no pre-processing, no regularization and no data augmentation except random color augmentations. Specifically, the backbone feature extractor is an S3D model [86] pre-trained on Kinetics 400 [35]. The output of each convolutional block is pooled, concatenated, and fed to a 2-layer MLP classification head. Such shortcut connections proved beneficial over a sequential model in early experiments, likely due to the multi-scale nature of manipulation artefacts. During training, the AdamW optimizer [45] minimizes a cross-entropy loss $\mathcal{L}_{\text{CE}}$ based on binary video labels until a validation loss stops improving. Additional details about hyperparameters and training can be found in Appendix C. For each model variant described below, Table 1 reports the average cross-entropy loss and AUROC over 3 runs on the official test split. While all models achieve satisfactory results on both datasets, performance drops when generalizing from DFDC to DFD. For this reason, we consider explanation metrics evaluated on DFDC more indicative of explanation quality in the following experiments.

Table 1: **Classification metrics:** L_{CE} is categorical cross-entropy (↓), A_{ROC} is the area under the receiver operating characteristic curve (↑). Average values over 3 runs, full results in Tables 3 and 4. Reported values account for class imbalance as detailed in Table 2.

	DFDC test		DFD	
	$\mathcal{L}_{\text{CE}}$	A_{ROC}	$\mathcal{L}_{\text{CE}}$	A_{ROC}
S3D Baseline	**.447**	**89.0**	.694	**80.2**
S3D Bilateral	.696	54.2	.746	45.8
S3D Gaussian	.542	81.8	.760	66.4
S3D TV Loss	**.460**	**87.4**	**.698**	75.8
S3D Cutout	.481	87.2	**.655**	**79.6**
MViT	**.430**	**96.4**	**.513**	**90.0**

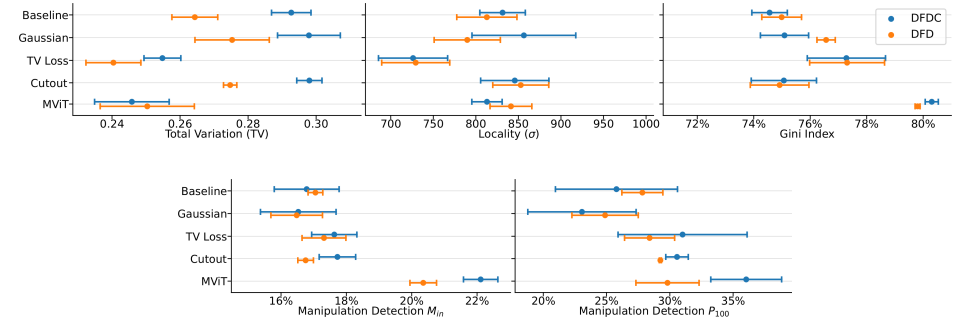


Figure 2: **Quantitative explanation metrics:** visual quality (top) and manipulation detection (bottom) for the evaluation subsets of DeepFake Detection Challenge (DFDC) and DeepFake Detection Dataset (DFD). Higher values indicate better explanation quality, except for TV and locality σ. Mean and standard deviation of 3 runs, full results in the appendix.

4.1 Quantitative results

Data preparation. As opposed to train-time augmentation, this term indicates transformations that are applied identically to all samples, *e.g.* face detection and cropping described above. For image DeepFakes, Wang *et al.* observe that generated images have a weaker high-frequency content than real ones [84] and the explanations of models that rely on this clue are dominated by uninterpretable high-frequency noise. They suggest pre-processing all samples with a bilateral filter [79] to encourage focussing on other more interpretable features. In the same spirit, we investigate whether removing high-frequency video components improves smoothness and locality of the explanations in a quantifiable way.

Two variants are considered: a per-frame bilateral filter [79] or a spatio-temporal Gaussian filter; both configured so that common artefacts remain visible. Only training videos are filtered, leaving validation, test and explanation splits unaltered. As reported in (Tab. 1), filtered videos result in lower classification performance, which corresponds to the observation in [84], and models trained with bilateral filtering fail to converge,thus we exclude them from explanation evaluation. Disappointingly, blurring does not improve explanation metrics in a consistent way (Fig. 2), except for a slightly higher Gini Index that indicates sparser heatmaps. It is surely possible that stronger filters could produce more marked effects, but at the cost of lower classification performance (Tab. 1). Otherwise, this outcome could be attributed to different generation techniques or compression formats between images and videos. Nevertheless, we recommend against this type of smoothing preprocessing [84] for video DeepFakes until proven more effective.

Regularization loss. Regularization refers to training-time techniques that smooth or constrain the loss landscape so that the optimization process yields more desirable solutions that generalize better and/or yield better explanations. A common technique is to add a per-layer Total Variation (TV) term to the loss function during training [66]. Considering the activation tensor $\boldsymbol{A}^\ell \in \mathbb{R}^{T\times H\times W}$ of an intermediate layer ℓ, its anisotropic total variation is:

$$\mathcal{L}^\ell_{\text{TV}} = \frac{1}{THW}\sum_d \Omega_{\text{1D}}(\boldsymbol{A}^\ell_d), \tag{5}$$

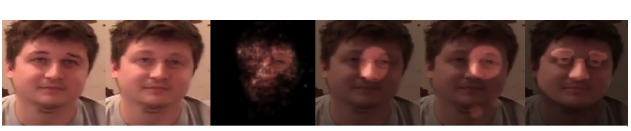
where the summation considers all 1D slices of $\boldsymbol{A}$ orthogonal to its axes and Ω_{1D} indicates the 1D total variation. Averaging over all convolutional blocks in our architecture, the optimization objective results $\mathcal{L} = \mathcal{L}_{\text{task}} + \alpha \mathbb{E}_\ell \left[\mathcal{L}_{\text{TV}}^\ell\right]$, where $\alpha \in \mathbb{R}^+$ is a hyperparameter. The additional term places a smoothness constrain on the activations of intermediate layers, which we hope will result in localized peaks in the heatmap corresponding to visible artefacts in the video, though TV does not control the location of such peaks.

The first effect of TV regularization is noticeable during the initial phases of training. For unconstrained models, we observe that $\mathbb{E}_\ell \left[\mathcal{L}_{\text{TV}}^\ell\right]$ tends to increase during the initial phase of training and stabilizes at around 0.5 after one epoch. On the other hand, when $\alpha = 1$, the optimization process is dominated by $\mathcal{L}_{\text{TV}}$ for the first epochs and classification loss starts decreasing only after this term drops below 0.1. From the results in Table 1, a strong TV regularization affects classification performance negatively. However, we also observe a significant improvement over the baseline for locality, sparsity, and manipulation in Figure 2. In fact, the average σ for DFDC decreases from 814 to 726, indicating more spatially-focused explanations. Also, the Gini Index increases from 75% to 77%, meaning that fewer pixels are responsible for the bulk of the heatmaps. With respect to manipulation detection, the heatmaps produced by TV-regularized models match more closely the ground-truth, resulting in higher P_{100} for both DFDC and DFD.

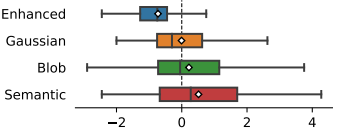
Video cutout. Cutout data augmentation which can greatly improve classification performance by masking input patches at random during training [20]. We adapt Cutout to video data by replacing masking with heavy spatio-temporal blur: since motion blur occurs naturally, the augmented samples are maintained closer to the data manifold. We expect Facial Cutout to guide the network towards more meaningful representations, where the relationship between semantic parts of the face are better understood, hence improving part-based manipulation detection. On the other hand, removing parts of the input might yield more spread out heatmaps, as the network learns to capture information from more diverse locations. In our experiments, we observe slightly better generalization to DFD for regularized models (Table 1), which confirms the regularization properties of Facial Cutout. However, the effects on explanations are limited, resulting in slightly higher Total Variation and manipulation detection scores (Figure 2).

Architecture. The architecture of a model represent a strong inductive biases on what features can be easily learned [26, 31]. As an alternative to the baseline S3D model, we consider another state-of-the-art architecture for video classification, namely a multi-scale vision transformer (MViT) [24]. The former, based on 3D convolutions, begins with forming local representations which are aggregated into more complex features in later layers. The latter, based on attention, allows all layers to attend to the input as a whole and encourages representation learning through progressive token aggregation. We expect the different inductive biases and information flow to affect the explanation heatmaps generated by these architectures. In particular, we fine-tune the MViT-B 16×4 variant with the default hyperparameters: random color augmentation, temporal subsampling, cosine learning rate annealing, and weight initialization from Kinetics 400. For ease of comparison, MViT explanations are obtained with SmoothGrad while attention-specific methods are left to future work.

For the classification task, MViT achieves the best performance on the two datasets (Tab. 1), which we attribute to the longer training cycle. The explanation heatmaps obtained with this architecture are also significantly smoother (TV) and sparser (Gini Index) than CNN-based



(a) From left to right: real video, fake video, enhanced heatmap, Gaussian matching, blob detection, semantic aggregation.



(b) User ratings for alternative explanation visualizations.

Figure 3: **User study.** *Left panel*: an example of the visualizations submitted to human observation for the study. *Right panel*: box plot of normalized score for each explanation visualization technique which semantic aggregation as preferred.

models, while spatial locality remains similar (σ). Furthermore, the bottom row of Fig. 2 indicates that MViT heatmaps are stronger detectors of manipulated areas, focusing most of the heatmap inside the ground-truth mask (M_{in}). We attribute these promising results to: i) a more robust classifier which can better distinguish fake videos and is thus likely to have learned a good representation of manipulation artefacts; and ii) the underlying inductive bias of attention and its effect on gradient propagation used for heatmap generation.

4.2 Communicating explanations

As discussed, gradient-based explanations often appear too noisy for users to easily parse. We propose four simple techniques to post-process heatmaps into increasingly more structured visualizations: i) **enhanced heatmaps**, clip extreme values and smooth to eliminate high-frequency noise; ii) **gaussian matching**, draw an ellipse corresponding to the mean and variance of each frame; iii) **blob detection**, run a DoG blob detector [16, 60] and highlight each blob according to its relevance; and iv) **semantic aggregation**, aggregate the heatmap into semantic regions and highlight each part based on its relevance.

A small-scale user study (34 participants) is carried out to quantify user satisfaction with respect to each of these visualization techniques. Each user is presented a set of 10 videos as in Fig. 3a and is asked to rate the four visualizations, which appear in random order. A score of 0 means that the visualization is not helpful to detect the DeepFake, while 5 means it easily allowed for its detection. To minimize appreciation bias, ratings are centered per-user before aggregation by subtracting the average score. From the results in Fig. 3b we observe a clear relationship between user satisfaction and more structured visualizations. However, when the classifier performs poorly and heatmaps are uninformative, users will be dissatisfied regardless of post-processing. However, we also note that when the classifier performs poorly, users will be generally dissatisfied.

5 Conclusion

The Explainable AI has developed a plethora of explanation methods of varying degrees of faithfulness. However, to the best of our knowledge, quantitative metrics to compare the quality of such explanations are lacking. This work attempts to lay out an objective evaluation framework for DeepFake explanations, which we hope will drive the development of detectors that are better aligned with human cognition. The main contribution of this paper is the introduction of a family of such metrics, novel or adapted from existing works, to measure visual quality and manipulation detection.

In our experiments we consider several techniques for training DeepFake detectors and study their impact on explainability metrics in a quantitative way, whereas previous work was limited to qualitative comparisons. We observe that TV regularization has the largest impact across most metrics. On the other hand, controlling high-frequency components of the input is of little utility, at least when realistic video compression settings are considered. Finally, we observe that recent architectures such as MViT significantly outperform any of the S3D variations in both detection and explanation quality. We recommend further study of transformer-based DeepFake classifiers and how to employ attention as an explanation.

Ethical statement. As DeepFake technology becomes increasingly accessible, so is the potential for malicious use. It is therefore urgent to present society with the necessary tools to address this problem and facilitate the safe and ethical use of these creations. We believe this line of work can bring positive societal impact by facilitating good governance and wider adoption of DeepFake detectors across all media. Furthermore, more explainable DeepFake detectors can be used to educate the public to better distinguish between real and generated content. From their perspective, users must feel confident about the technologies that routinely affects their interactions, which may otherwise fall victim to mistrust.

Limitations and future work. This project leads to many natural avenues for future research in Explainable AI. First, although the proposed metrics are drawn from existing literature and are based on sound mathematical foundations, an extensive study of the correlation between these metrics and human preference would increase their reliability. Second, it is surely possible to conceive more refined metrics for DeepFake detection to address the shortcomings discussed in Section 3.2. For instance, we have already mentioned that locality (σ) favors unimodal over multimodal heatmaps, whereas more faceted metrics of localization are desirable. Third, as made evident from the experiments on DeepFake Dataset, when classification performance is not perfect explanations can be meaningless. Thus, combining explanations and uncertainty estimation would provide a more complete picture of any DeepFake detector. Finally, we remark that the proposed metrics are not meant to supplant human judgment, *e.g.* user studies, but rather to provide a non-interactive and repeatable benchmark that is more suitable for guiding the development and facilitating the deployment of better DeepFake detectors.

References

[1] Samira Abnar and Willem Zuidema. Quantifying Attention Flow in Transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.385.

[2] Darius Afchar, Vincent Nozick, Junichi Yamagishi, and Isao Echizen. Mesonet: A compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[3] Shruti Agarwal, Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. Protecting World Leaders Against Deep Fakes. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, volume 1, 2019.

[4] David Alvarez Melis and Tommi Jaakkola. Towards Robust Interpretability with Self-Explaining Neural Networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[5] Irene Amerini, Leonardo Galteri, Roberto Caldelli, and Alberto Del Bimbo. Deepfake Video Detection through Optical Flow Based CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[6] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for Deep Neural Networks. In *International Conference on Learning Representations*, February 2018.

[7] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A Video Vision Transformer. *arXiv:2103.15691 [cs]*, March 2021.

[8] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):e0130140, July 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140.

[9] Federico Baldassarre, Kevin Smith, Josephine Sullivan, and Hossein Azizpour. Explanation-Based Weakly-Supervised Learning of Visual Relations with Graph Networks. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *ECCV 2020*, Lecture Notes in Computer Science, pages 612–630. Springer International Publishing, 2020. ISBN 978-3-030-58604-1. doi: 10.1007/978-3-030-58604-1_37.

[10] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Richard Benjamins, Raja Chatila, and Francisco Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, June 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012.

[11] Luca Bondi, Edoardo Daniele Cannas, Paolo Bestagini, and Stefano Tubaro. Training Strategies and Data Augmentations in CNN-based DeepFake Video Detection. In *2020 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6, December 2020. doi: 10.1109/WIFS49906.2020.9360901.

[12] Nicolò Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video Face Manipulation Detection Through Ensemble of CNNs. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5012–5019, January 2021. doi: 10.1109/ICPR48806.2021.9412711.

[13] Chunshui Cao, Xianming Liu, Yi Yang, Yinan Yu, Jiang Wang, Zilei Wang, Yongzhen Huang, Liang Wang, Chang Huang, Wei Xu, Deva Ramanan, and Thomas S. Huang. Look and Think Twice: Capturing Top-Down Visual Attention with Feedback Convolutional Neural Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2956–2964, December 2015. doi: 10.1109/ICCV.2015.338.

[14] Hila Chefer, Shir Gur, and Lior Wolf. Generic Attention-Model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *arXiv:2103.15679 [Cs]*, March 2021.

[15] Davide Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining EfficientNet and Vision Transformers for Video Deepfake Detection. *arXiv:2107.02612 [cs]*, July 2021.

[16] Robert T. Collins. Mean-shift blob tracking through scale space. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings.*, volume 2, pages II–234. IEEE, 2003.

[17] Jesse Damiani. A Voice Deepfake Was Used To Scam A CEO Out Of $243,000. https://www.forbes.com/sites/jessedamiani/2019/09/03/a-voice-deepfake-was-used-to-scam-a-ceo-out-of-243000/, September 2019.

[18] Hao Dang, Feng Liu, Joel Stehouwer, Xiaoming Liu, and Anil K. Jain. On the detection of digital face manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5781–5790, 2020.

[19] Deepfakes Team. Deepfakes, November 2021.

[20] Terrance DeVries and Graham W. Taylor. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv:1708.04552 [cs]*, November 2017.

[21] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The DeepFake Detection Challenge (DFDC) Dataset. *arXiv:2006.07397 [cs]*, October 2020.

[22] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *arXiv:2010.11929 [Cs]*, September 2020.

[23] European Union. General Data Protection Regulation (GDPR). https://gdpr.eu/, 2018.

[24] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. *arXiv:2104.11227 [cs]*, April 2021.

[25] Ruth C. Fong and Andrea Vedaldi. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3449–3457, Venice, October 2017. IEEE. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.371.

[26] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*, September 2018.

[27] Generated Photos Team. Generated Photos Dataset. https: //generated.photos/, 2018.

[28] Leilani H. Gilpin, David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining Explanations: An Overview of Interpretability of Machine Learning. *The 5th IEEE International Conference on Data Science and Advanced Analytics (DSAA 2018).*, June 2018.

[29] Corrado Gini. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. Tipogr. di P. Cuppini, 1912.

[30] David Güera and Edward J. Delp. Deepfake Video Detection Using Recurrent Neural Networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6, November 2018. doi: 10.1109/AVSS.2018.8639163.

[31] Niall Hurley and Scott Rickard. Comparing Measures of Sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, October 2009. ISSN 1557-9654. doi: 10.1109/TIT.2009.2027527.

[32] Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1357.

[33] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. XRAI: Better Attributions Through Regions. *arXiv:1906.02825 [cs, stat]*, August 2019.

[34] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[35] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics Human Action Video Dataset. *arXiv:1705.06950 [cs]*, May 2017.

[36] Minha Kim, Shahroz Tariq, and Simon S. Woo. FReTAL: Generalizing Deepfake Detection Using Knowledge Distillation and Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1012, 2021.

[37] Pieter-Jan Kindermans, Kristof Schütt, Klaus-Robert Müller, and Sven Dähne. Investigating the influence of noise and distractors on the interpretation of neural networks. In *NIPS 2016 Workshop on Interpretable Machine Learning in Complex Systems*, November 2016.

[38] Kurt Koffka. *Principles of Gestalt Psychology*. Routledge, 2013.

[39] Alexander Kolesnikov and Christoph H. Lampert. Seed, Expand and Constrain: Three Principles for Weakly-Supervised Image Segmentation. In *ECCV (4)*, January 2016.

[40] Marek Kowalski. FaceSwap, October 2021.

[41] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Advancing High Fidelity Identity Swapping for Forgery Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5074–5083, 2020.

[42] Yuezun Li and Siwei Lyu. Exposing deepfake videos by detecting face warping artifacts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

[43] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[44] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3207–3216, 2020.

[45] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR 2019*, 2019.

[46] Scott M Lundberg and Su-In Lee. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

[47] M. Masood, Marriam Nawaz, K. M. Malik, A. Javed, and Aun Irtaza. Deepfakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward. *ArXiv*, 2021.

[48] Falko Matern, Christian Riess, and Marc Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92. IEEE, 2019.

[49] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv:2012.10289 [cs]*, December 2020.

[50] Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, February 2019. ISSN 0004-3702. doi: 10.1016/j.artint.2018.07.007.

[51] Sina Mohseni, Niloofar Zarei, and Eric D. Ragan. A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3-4):24:1–24:45, August 2021. ISSN 2160-6455. doi: 10.1145/3387166.

[52] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65:211–222, May 2017. ISSN 0031-3203. doi: 10.1016/j.patcog.2016.11.008.

[53] Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, 73:1–15, February 2018. ISSN 1051-2004. doi: 10.1016/j.dsp.2017.10.011.

[54] Joao C. Neves, Ruben Tolosana, Ruben Vera-Rodriguez, Vasco Lopes, Hugo Proença, and Julian Fierrez. GANprintR: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14 (5):1038–1048, 2020.

[55] Huy H. Nguyen, Fuming Fang, Junichi Yamagishi, and Isao Echizen. Multi-task Learning for Detecting and Segmenting Manipulated Facial Images and Videos. In *2019 IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–8. IEEE, 2019.

[56] Huy H. Nguyen, Junichi Yamagishi, and Isao Echizen. Capsule-forensics: Using Capsule Networks to Detect Forged Images and Videos. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311, May 2019. doi: 10.1109/ICASSP.2019.8682602.

[57] Nick Dufour and Andrew Gully. Deep Fake Detection Dataset by Google and JigSaw, 2019.

[58] Y. Nirkin, I. Masi, A. Tran, Tal Hassner, and G. Medioni. On Face Segmentation, Face Swapping, and Face Perception. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, 2018. doi: 10.1109/FG.2018.00024.

[59] Yuval Nirkin, Yosi Keller, and Tal Hassner. FSGAN: Subject Agnostic Face Swapping and Reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7184–7193, 2019.

[60] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011. ISSN 1533-7928.

[61] Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*, September 2018.

[62] Samuele Pino, Mark James Carman, and Paolo Bestagini. What's wrong with this video? Comparing Explainers for Deepfake Detection. *arXiv:2105.05902 [cs]*, May 2021.

[63] Jiameng Pu, Neal Mangaokar, Lauren Kelly, Parantapa Bhattacharya, Kavya Sundaram, Mobin Javed, Bolun Wang, and Bimal Viswanath. Deepfake Videos in the Wild: Analysis and Detection. *arXiv:2103.04263 [cs]*, March 2021.

[64] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. *arXiv:1602.04938 [cs, stat]*, August 2016.

[65] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Niessner. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[66] Leonid I. Rudin, Stanley Osher, and Emad Fatemi. Nonlinear total variation based noise removal algorithms. *Physica D: nonlinear phenomena*, 60(1-4):259–268, 1992.

[67] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the Visualization of What a Deep Neural Network Has Learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11): 2660–2673, November 2017. ISSN 2162-2388. doi: 10.1109/TNNLS.2016.2599820.

[68] Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. *Proceedings of the IEEE*, 109(3):247–278, March 2021. ISSN 1558-2256. doi: 10.1109/JPROC.2021.3060483.

[69] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization. *International Journal of Computer Vision*, 128(2):336–359, February 2020. ISSN 0920-5691, 1573-1405. doi: 10.1007/s11263-019-01228-7.

[70] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, July 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x.

[71] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv:1312.6034 [cs]*, April 2014.

[72] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. SmoothGrad: Removing noise by adding noise. *arXiv:1706.03825 [cs, stat]*, June 2017.

[73] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin A. Riedmiller. Striving for Simplicity: The All Convolutional Net. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Workshop Track Proceedings*, 2015.

[74] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. *arXiv:1703.01365 [cs]*, June 2017.

[75] Supasorn Suwajanakorn, Steven M. Seitz, and Ira Kemelmacher-Shlizerman. Synthesizing Obama: Learning lip sync from audio. *ACM Transactions on Graphics*, 36(4): 95:1–95:13, July 2017. ISSN 0730-0301. doi: 10.1145/3072959.3073640.

[76] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Niessner. Face2Face: Real-Time Face Capture and Reenactment of RGB Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2387–2395, 2016.

[77] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics*, 38(4):66:1–66:12, July 2019. ISSN 0730-0301. doi: 10.1145/3306346.3323035.

[78] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, December 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2020.06.014.

[79] Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*, pages 839–846. IEEE, 1998.

[80] Loc Trinh, Michael Tsang, Sirisha Rambhatla, and Yan Liu. Interpretable and Trustworthy Deepfake Detection via Dynamic Prototypes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1973–1983, 2021.

[81] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L. Griffiths. Are Convolutional Neural Networks or Transformers more like human vision? In *Cognitive Science Society*, July 2021.

[82] Cristian Vaccari and Andrew Chadwick. Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media + Society*, 6(1):2056305120903408, January 2020. ISSN 2056-3051. doi: 10.1177/2056305120903408.

[83] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NIPS*, December 2017.

[84] Kaili Wang, Jose Oramas, and Tinne Tuytelaars. Towards Human-Understandable Visual Explanations: Imperceptible High-frequency Cues Can Better Be Removed. *arXiv:2104.07954 [cs]*, April 2021.

[85] Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 2019. ISSN 1927-0321. doi: 10.22215/timreview/1282.

[86] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 305–321, 2018.

[87] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*, pages 2048–2057. PMLR, June 2015.

[88] Shawn Xu, Subhashini Venugopalan, and Mukund Sundararajan. Attribution in Scale and Space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9680–9689, 2020.

[89] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.

[90] Changqian Yu, Changxin Gao, Jingbo Wang, Gang Yu, Chunhua Shen, and Nong Sang. Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation. *International Journal of Computer Vision*, pages 1–18, 2021.

[91] Matthew D. Zeiler and Rob Fergus. Visualizing and Understanding Convolutional Networks. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 818–833, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10590-1. doi: 10.1007/978-3-319-10590-1_53.

[92] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, October 2016. ISSN 1558-2361. doi: 10.1109/LSP.2016.2603342.

[93] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning Deep Features for Discriminative Localization. *arXiv:1512.04150 [cs]*, December 2015.

[94] Hang Zhou, Yu Liu, Ziwei Liu, Ping Luo, and Xiaogang Wang. Talking face generation by adversarially disentangled audio-visual representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9299–9306, 2019.

[95] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. WildDeepfake: A Challenging Real-World Dataset for Deepfake Detection. *arXiv:2101.01456 [cs]*, January 2021.