Explaining Deepfake Detection by Analysing Image Matching

Shichao Dong[1,*], Jin Wang[1,*], Jiajun Liang[1], Haoqiang Fan[1], and Renhe Ji[1,†]

[1]MEGVII Technology
{dongshichao,wangjin,liangjiajun,fhq,jirenhe}@megvii.com

Abstract. This paper aims to interpret how deepfake detection models learn artifact features of images when just supervised by binary labels. To this end, three hypotheses from the perspective of image matching are proposed as follows. 1. Deepfake detection models indicate real/fake images based on visual concepts that are neither source-relevant nor target-relevant, that is, considering such visual concepts as artifact-relevant. 2. Besides the supervision of binary labels, deepfake detection models implicitly learn artifact-relevant visual concepts through the FST-Matching (*i.e.* the matching **f**ake, **s**ource, **t**arget images) in the training set. 3. Implicitly learned artifact visual concepts through the FST-Matching in the raw training set are vulnerable to video compression. In experiments, the above hypotheses are verified among various DNNs. Furthermore, based on this understanding, we propose the FST-Matching Deepfake Detection Model to boost the performance of forgery detection on compressed videos. Experiment results show that our method achieves great performance, especially on highly-compressed (*e.g.* c40) videos.

Keywords: deepfake detection, image matching, interpretability.

1 Introduction

Recently, deepfake methods [11, 23, 21, 40, 39] have exhibited superior performance in synthesizing realistic faces. Such face forgeries may easily be used by attackers for malicious purposes, causing severe social problems and political threats. To this end, plenty of studies [32, 1] have achieved great success in detecting various manipulated media by simply considering it as a binary classification task. However, understanding how these models learn artifact features of images when just supervised by binary labels (real/fake) is still a challenge to state-of-the-art algorithms.

In this paper, we aim to interpret the success of deepfake detection models from the novel perspective of image matching. We consider the matching images as follows. As shown in Fig 1, the face of the source image is manipulated with representations of the target image to generate the corresponding fake image.

* Equal contribution
† Corresponding author

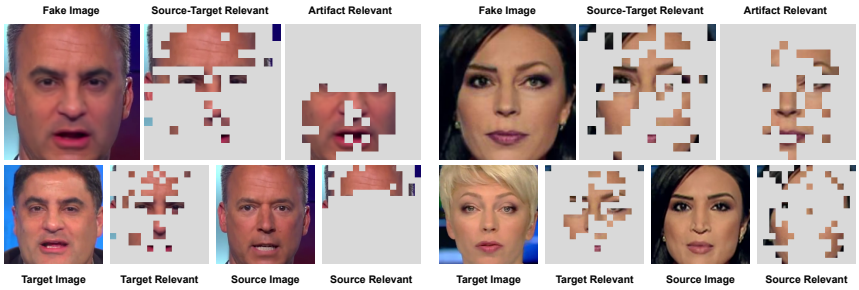

Fig. 1. The relationship between source/target-relevant visual concepts and artifact-relevant visual concepts. Here, visual concepts represent image regions such as eyes, mouths and foreheads of human faces. In this paper, we find that well-trained deepfake detection models mainly consider artifact-relevant visual concepts as neither source-relevant nor target-relevant from the perspective of image matching.

Then the above **f**ake image, **s**ource image and **t**arget image are considered as the matching images, termed as the FST-Matching. To this end, we design different metrics to quantitatively evaluate the effectiveness of image matching and propose three hypotheses as follows.

Hypothesis 1: Deepfake detection models indicate real/fake images based on visual concepts that are neither source-relevant nor target-relevant, that is, considering such visual concepts as artifact-relevant. In this paper, visual concepts represent the image regions such as the mouths, noses or eyes of human faces. Intuitively, fake images are generated from visual concepts that are either from source images or target images. However, some visual concepts may inevitably be manipulated by deepfake methods, causing them to be different from both source images and target images. Well-trained deepfake detection models are supposed to indicate real/fake images based on both source-irrelevant and target-irrelevant visual concepts.

Hypothesis 2: Besides the supervision of binary labels, deepfake detection models implicitly learn artifact-relevant visual concepts through the FST-Matching in the training set. Intuitively, binary labels are not sufficient enough to accomplish the deepfake detection task. Training images usually contain other artifact-irrelevant visual concepts, such as the identity of images. Such visual concepts may co-appear on certain real/fake images, causing deepfake detection models to learn biased representations of the forgeries. For example, deepfake detection models may infer the results based on the gender of images if real images are all male and fake images are all female. To this end, FST-Matching images are supposed to help deepfake detection models to discard artifact-irrelevant visual concepts and focus on artifact-relevant visual concepts, since they share common artifact-irrelevant visual concepts but are annotated with opposite labels.

Hypothesis 3: Implicitly learned artifact visual concepts through the FST-Matching in the raw training set are vulnerable to the video compression. Deepfake detection models trained on raw images usually suffer from significant performance drop when testing on compressed images [24, 50, 32]. We assume that it is because the implicit learning of artifact visual concepts through FST-Matching is fragile to the video compression. Specifically, the implicitly learned artifact visual concepts may become indistinguishable from compressed source visual concepts and target visual concepts on fake images due to the compression, causing deepfake detection models to make false predictions.

Methods: To verify the proposed hypotheses, we propose an explanation method based on the Shapley value [35] to interpret the predictions of deepfake detection models with various backbones. The Shapley value was firstly proposed in game theory [35] and is widely used in recent studies [27, 2] to interpret the representations inside DNNs. Specifically, the Shapley value unbiasedly estimates the contributions of each player to the total award of the game. It naturally satisfies four properties, *i.e.* the linearity property, the dummy property, the symmetry property, and the efficiency property [41], which ensures its fairness and trustworthiness. Based on the Shapley value, we evaluate the visual concepts on images from the novel perspective of image matching to verify the proposed hypotheses.

Furthermore, during the verification of hypotheses, we surprisingly find the learned source/target visual concepts are more consistent among compressed images than the implicitly learned artifact visual concepts on images. Combined with the understanding of hypothesis 1, we then devise a simple model by disentangling source/target-irrelevant representations from the source/target visual concepts to indicate images (termed as the FST-Matching Deepfake Detection Model), which aims to boost the performance of the forgery detection on compressed videos. Results in our experiments show that such simple architecture achieves great performance, especially on highly compressed (*e.g.* c40) videos.

Contributions: Our contributions can be summarized as follows.

1. We propose a method to interpret the success of deepfake detection models from the novel perspective of image matching, *i.e.* the FST-Matching.

2. Three hypotheses from the perspective of the FST-Matching are proposed and verified, which offers new insights into the task of deepfake detection.

3. We further propose the FST-Matching Deepfake Detection Model to improve the performance on compressed videos.

2 Related work

2.1 Deepfake detection

The goal for deepfake detection is to classify the input media as either real or fake. Previous studies of deepfake detection mainly focused on improving the model performance on various datasets. Some methods [1, 3, 8, 30–32] considered it as a binary classification task and directly trained models on the largely-collected dataset, such as Celeb-DF [25], DFDC [9], FF++ [32] and *etc.* These

methods achieved great performance on the in-dataset evaluation, *i.e.* testing models on images manipulated by learned deepfake methods. However, these methods often failed to detect unseen datasets with newly proposed deepfake methods. To this end, other studies [51, 49, 54, 20] aim to increase the generalization of deepfake detection models. These methods usually assumed that fake images share common human-perceived artifact representations introduced in the process of deepfake methods, such as blending boundaries [24], geometric features [37] and frequency features [26, 28, 22, 14]. However, such assumptions usually represent human's understanding of artifact representations and may not hold in all real-life scenarios. It still presents continuous challenges to correctly understand the key differences between real and fake images, *i.e.* exploring the essence of the artifact representations on images.

To the best of our knowledge, studies focused on interpreting the learned representations of deepfake detection models are rare. In this paper, we aim to interpret deepfake detection models from the novel perspective of image matching to demonstrate what artifact representations are to deepfake detection models, how they learned artifact representations and how to further boost their performance in real-life scenarios.

2.2 Interpretability of DNNs

Previous studies on the interpretability of DNNs can be roughly divided into two categories. Some studies [43, 29, 10, 36, 42, 53] focused on semantic explanations for DNNs by visualizing the learned visual concepts. Grad-CAM [34] and Grad-CAM++ [5] explored the attribution maps of input images based on gradient information. Zhou *et. al.* [52] visualized the actual receptive fields of various units inside the DNNs. Fong *et. al.* [12] explored the relationship between multiple filters and learned semantic visual concepts. Zhang *et. al.* proposed to explore the relationships between the learned semantic visual concepts of DNNs via a graph model[47] and a decision tree [48]. However, different from general classification tasks, deepfake detection models aim to learn artifact-relevant visual concepts on images. Such representation is often imperceptible to people, making it difficult to evaluate the correctness of the explanation results derived from the above methods. Moreover, other studies proposed to explain the representations of DNNs mathematically to refrain from human evaluation of semantic representation. To this end, some studies proposed to understand DNNs based on entropy-based methods [15, 7]. Some studies explored the representations of DNNs from a game-theoretical view [45, 44, 46]. However, although the above methods can be theoretically applied to various types of DNNs, it still remains a challenge to further exploit the explanation results to instruct the learning of specific tasks, such as deepfake detection.

In this paper, we aim to bridge the gap between the general explanation results and learning better deepfake detection models from the novel perspective of image matching. To this end, we designed the FST-Matching Deepfake Detection Model based on our explanation results and further boosted the performance on compressed videos.

3 Algorithms

In this section, given a well-trained deepfake detection model, we aim to interpret its prediction from the novel perspective of image matching. To this end, three hypotheses are proposed. To verify these hypotheses, we propose an explanation method to evaluate the contributions of visual concepts on images based on the Shapley value [35]. Please see supplementary materials for more information about the Shapley value.

3.1 Artifact representations for deepfake detection models

Hypothesis 1 : Deepfake detection models indicate real/fake images based on visual concepts that are neither source-relevant nor target-relevant, that is, considering such visual concepts as artifact-relevant.

In this section, given a well-trained deepfake detection model $v_d(\cdot)$ (also termed as the detection encoder in this paper), we aim to evaluate the learned visual concepts on input images from the perspective of image matching. Specifically, we aim to explore what visual concepts on input images are considered as source-relevant, target-relevant and artifact-relevant. Then, we expect to evaluate the relationship between these visual concepts to verify the hypothesis.

The core challenge is to decide fairly what visual concepts are related to the source, target and artifact representations. Specifically, we do not annotate these visual concepts on images manually since it usually represents human's understanding of artifact representations, rather than the artifact representations inside the models. To this end, we train a source encoder $v_s(\cdot)$ and a target encoder $v_t(\cdot)$ to indicate the source/target-relevant visual concepts on images.

Intuitively, each fake image shares certain common visual concepts with its corresponding source and target image. We believe that when the source encoder v_s classifies each fake image and its corresponding source image as the same category, v_s would tend to focus on source-relevant visual concepts on each fake image. The same way goes for the target encoder v_t. Specifically, we use the additional attribute labels[1] of images to train v_s and v_t for convenience. To train the source/target encoder v_s/v_t, each fake image is considered as the same attribute label as the corresponding source/target image. Each real image is considered as its original attribute label.

We use the Shapley value [35] to evaluate the regional contributions of visual concepts on images to the prediction of each encoder. To reduce the computation cost, we divide the input image into $L \times L$ grids and calculate the contribution of each grid respectively. Let $G = \{g_{11}, g_{12}, ..., g_{LL}\}$ denote the set of all grids. $\phi_{v_d} \in R^{L\times L}, \phi_{v_s} \in R^{L\times L}, \phi_{v_t} \in R^{L\times L}$ represent the contributions of all grids to the prediction of the detection encoder v_d, the source encoder v_s and the target encoder v_t respectively. In this way, ϕ_{v_d}, ϕ_{v_s} and ϕ_{v_t} indicate the artifact, source and target visual concepts on images respectively. More specifically, given

[1] implemented as the identity labels of images for convenience.

$\forall g_{ij} \in G$, it is considered to be artifact-relevant if $\phi_{v_d}(g_{ij}|G) > 0$ and artifact-irrelevant if $\phi_{v_d}(g_{ij}|G) \leq 0$. The same way goes for the source encoder v_s and target encoder v_t.

Based on the grid-level contributions, we propose a metric to evaluate the relationship between the artifact-relevant visual concepts, source-relevant visual concepts and target-relevant visual concepts. According to the hypothesis, deepfake detection models are supposed to consider artifact-relevant visual concepts as neither source-relevant nor target-relevant. Therefore, artifact-relevant visual concepts are supposed to barely have intersections with source/target-relevant visual concepts. To this end, we firstly generate a mask $M_\tau = I(max(\phi_{v_s}, \phi_{v_t}) > \tau)$ to denote the most source/target-relevant visual concepts, where $I(\cdot)$ is the indicator function and τ is a certain threshold. $I(\cdot)$ returns 1 if the condition inside is valid, otherwise $I(\cdot)$ returns 0. The metric is then designed to evaluate the intensities of the intersections between these visual concepts as follows.

$$Q_\tau = \frac{(1 - M_\tau) \cdot \phi_{v_d}}{\sum_{g_{ij} \in G} [1 - M_\tau(g_{ij})]} - \frac{M_\tau \cdot \phi_{v_d}}{\sum_{g_{ij} \in G} M_\tau(g_{ij})} \tag{1}$$

where $\cdot$ denotes the inner product. The first term measures the average intensities of the intersections between source/target-irrelevant visual concepts and artifact-relevant visual concepts. The second term measures the average intensities of the intersections between the source/target-relevant visual concepts and artifact-relevant visual concepts. $Q_\tau > 0$ represents that artifact-relevant visual concepts are more related to source/target-irrelevant visual concepts than the source/target-relevant visual concepts. $Q_\tau < 0$ represents that artifact-relevant visual concepts are less related to source/target-irrelevant visual concepts than the source/target-relevant visual concepts.

3.2 Learning the artifact representations

Hypothesis 2: Besides the supervision of binary labels, deepfake detection models implicitly learn artifact-relevant visual concepts through the FST-Matching in the training set.

In this section, to verify the hypothesis, we expect to evaluate how the FST-Matching in the training set affects the learning of deepfake detection models. Specifically, FST-Matching in the training set means that real images contain the corresponding source and target images of fake images. To this end, we train two models with the paired training set and the unpaired training set separately. In the paired training set, the real images are only the corresponding source images and target images of fake images. In the unpaired images, the real images are of the same number as real images in the paired training set but do not correspond to any fake images. Then we compare the ACC, video-level AUC and the proposed metric Q_τ on these two models to evaluate the effectiveness of the FST-Matching.

3.3 Vulnerability of artifact representations to video compression

Hypothesis 3: Implicitly learned artifact visual concepts through the FST-Matching in the raw training set are vulnerable to the video compression.

In this section, to verify the hypothesis, we aim to measure the stability of implicitly learned artifact visual concepts to the video compression. Note that the detection encoder v_d is firstly trained on raw images and tested on compressed images afterwards. To this end, we design the stability metric to evaluate the changes among artifact visual concepts under the conditions of different compression rates *i.e.* c23, c40. The stability metric is designed as follows.

$$\delta_{v_d} = E_{cmp \in \{c23, c40\}}[cos(\phi_{v_d}^{cmp}, \phi_{v_d}^{raw})] \tag{2}$$

where $\phi_{v_d}^{cmp}$ represents the grids contributions to the predictions of the detection encoder v_d when tested on the compressed images. $\phi_{v_d}^{raw}$ represents the grids contributions tested on the raw images. $cos(\cdot, \cdot)$ denotes the operation of calculating the cosine similarity. A smaller value of $\delta_{v_d} \in [-1, 1]$ indicates that the implicitly learned artifact visual concepts are vulnerable to the compression. Moreover, we also evaluate the stability of the learned source/target visual concepts for source/target encoder v_s/v_t on compressed videos for more comparisons.

3.4 FST-Matching Deepfake Detection Model

Based on the understanding of deepfake detection models from the perspective of FST-Matching, we propose the FST-Matching Deepfake Detection Model to further boost the performance of deepfake detection models on compressed videos. During the verification of hypothesis 3, we surprisingly found that source/target visual concepts learned by the source encoder v_s and the target encoder v_t (*i.e.* ϕ_{v_s} and ϕ_{v_t}) are more consistent than the artifact visual concepts implicitly learned by the detection encoder v_d (*i.e.* ϕ_{v_d}) on compressed images (shown in the bottom of Fig 2). Inspired by the understanding of hypothesis 1, we believe that directly disentangling source/target-irrelevant representations from source/target visual concepts to indicate images may improve the model performance on compressed videos. Please see supplementary for detailed verification.

The structure of the FST-Matching Deepfake Detection Model is shown in Fig. 2, which aims to classify face forgeries based on source/target-irrelevant visual concepts on images according to hypothesis 1. To this end, we first use the Source Feature Encoder and the Target Feature Encoder to directly learn the source feature $f_s \in R^{B \times C_s}$ and the target feature $f_t \in R^{B \times C_t}$ on images. B indicates the number of input images. C_s and C_t indicate the number of output channels. Then we design the Feature Disentanglement Module to automatically disentangle the source/target-irrelevant feature f_s^{ir}, f_t^{ir} and source/target-relevant feature f_s^r, f_t^r on the channel-level. Similar to [19], we use the channel-wise attention vectors $a_s \in R^{B \times C_s}$ and $a_t \in R^{B \times C_t}$ to disentangle f_s and f_t, which are calculated as follows.

$$a_s = \sigma(MLP(f_s)),\ a_t = \sigma(MLP(f_t)) \tag{3}$$

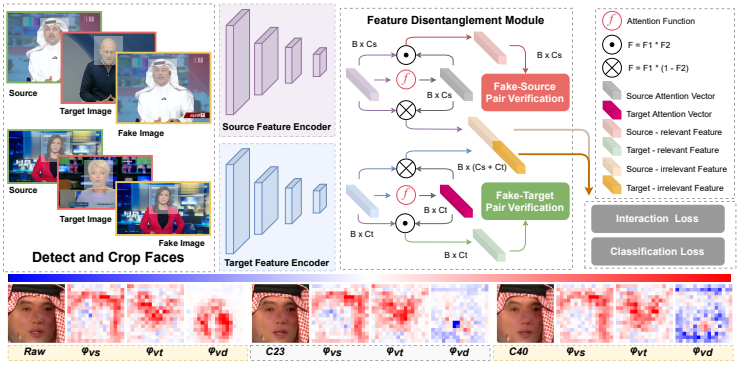

Fig. 2. The FST-Matching Deepfake Detection Model. As shown in the bottom of the figure, we surprisingly find that ϕ_{v_s} and ϕ_{v_t} are more robust to video compression than ϕ_{v_d}. To this end, we use a Source Feature Encoder and a Target Feature Encoder to explicitly learn the source and target representations on images. The Feature Disentanglement Module further extracts source/target-irrelevant representations to indicate the realism of images *i.e.* real or fake.

where MLP denotes the multi-layer perceptron and σ denotes the sigmoid function. In this way, the source and target relevant feature f_s^r, f_t^r are calculated as $f_s^r = a_s \circ f_s$ and $f_t^r = a_t \circ f_t$. The source and target irrelevant feature f_s^{ir}, f_t^{ir} are calculated as $f_s^{ir} = (1 - a_s) \circ f_s$ and $f_t^{ir} = (1 - a_t) \circ f_t$. Here $\circ$ denotes the channel-wise product.

To ensure the effectiveness of the feature disentanglement, we use the Fake-Source Pair Verification module to classify f_s^r as the same attribute label of the source images[2]. Similarly, f_t^r is classified as the same attribute label of the target image through the Fake-Target Pair Verification module. f_s^{ir} and f_t^{ir} are then concatenated to predict the final real/fake label of the input image. Let y_s, y_t, y_d denote the source attribute label, target attribute label and forgery detection label of the image. $\hat{y}_s, \hat{y}_t, \hat{y}_d$ denote the predicted source attribute, target attribute and forgery prediction. The classification loss of the FST-Matching Deepfake Detection Model is designed as follows.

$$Loss_{cls} = -E[y_d log\hat{y}_d] - \lambda_s E[y_s log\hat{y}_s] - \lambda_t E[y_t log\hat{y}_t] \tag{4}$$

Moreover, inspired by [45], we design another loss to further strengthen the interaction between f_s^{ir} and f_t^{ir} for the final prediction. Let $h(\cdot)$ denote the final prediction module. The interaction loss aims to increase the additional award caused by the coalition $[f_s^{ir}, f_t^{ir}]$ *w.r.t.* the sum of the award when f_s^{ir} and f_t^{ir} contribute to the final prediction individually. The interaction loss is designed as follows.

$$Loss_{interaction} = -E[h([f_s^{ir}, f_t^{ir}]) - h([\mathbf{0}, f_t^{ir}]) - h([f_s^{ir}, \mathbf{0}]) + h([\mathbf{0}, \mathbf{0}])] \tag{5}$$

[2] implemented as the identity labels of images for convenience.

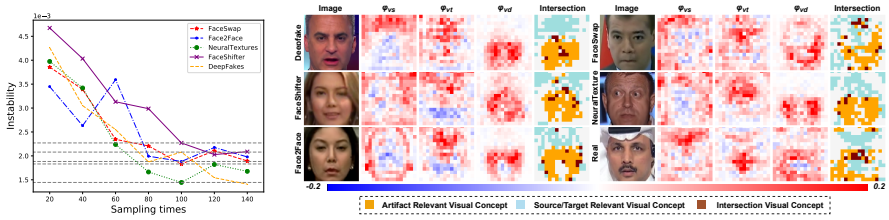

Fig. 3. Instability of the Shapley value (left) and verification of hypothesis 1(right). The left figure shows that as the sampling times increase, the Shapley value becomes stable. The right figure shows the visualization of source, target and artifact visual concepts, *i.e.* ϕ_{v_s}, ϕ_{v_t} and ϕ_{v_d}. Results show that artifact-relevant visual concepts barely have intersections with source/target-relevant visual concepts among various manipulation algorithms, which supports hypothesis 1.

where $\mathbf{0}$ represents the zero vector in the same size with f_s^{ir} and f_t^{ir}. $h([\mathbf{0}, \mathbf{0}]$ represents the basic score when neither f_s^{ir} nor f_t^{ir} contributes to the final prediction. The overall loss is designed as follows.

$$Loss = Loss_{cls} + \lambda_{inter} Loss_{interaction} \tag{6}$$

4 Experiment

4.1 Implementation details

DNNs & Datasets: To verify the proposed hypotheses, we conduct various experiments on different backbones. Specifically, we used ResNet-18/34 [18] and EfficientNet-b3 [38] as the backbones for the detection encoder v_d, v_s and v_t. Besides, we also used the pre-trained models released in [32] and [49] for the detection encoder v_d for more comparisons with state-of-the-art methods.

We trained and tested our models on the widely-used FF++ [32] dataset. FF++ [32] dataset contains 5000 videos, including 1000 original videos and 4000 fake videos manipulated by different forgery methods, such as Deepfake [11], FaceSwap [21], FaceShifter [23], NeuralTextures [39] and Face2Face [40]. All models were pre-trained on the ImageNet [33] dataset and fine-tuned on FF++ [32]. Moreover, the attribute label of the input image is set as the identity of the image for convenience. Specifically, for the fake image, the source/target encoder is expected to classify the image as the identity of its corresponding source/target image. For the real image, the source encoder and the target encoder are both expected to classify the image as its own original identity.

Implementation of the Shapley value: The precise calculation of the Shapley value is computationally intolerable. To this end, we used the sampling-based method [4] to approximately calculate the contributions of all the visual concepts. During the sampling process, the unsampled grids of images were set as the baseline value, which is set to be zero in this paper. Moreover, we used the

selected scalar before the softmax layer corresponding to the ground truth label of the image as the output score for all the encoders.

4.2 Fairness of the Shapley value

Accuracy of the Shapley value To ensure the stability of the approximated Shapley value, we evaluated the effect of sampling times T *w.r.t* the change of the Shapley value. Specifically, similar to [44], we repeated the sampling procedures [4] two times for the same sampling times T to get ϕ_1 and ϕ_2 respectively. Then we measured the change between ϕ_1 and ϕ_2 *w.r.t* to the sampling times T via the instability metric $\frac{||\phi_1-\phi_2||_2}{||\phi_1+\phi_2||_2}$ among all test images. As shown in Fig 3, we calculated the instability metric for ResNet18-based ϕ_{v_d} for different samping times. Results show that when $T \geq 100$, we get the relatively stable Shapley value, which ensures the fairness of our results.

4.3 Verification of hypotheses

Verification of hypothesis 1. Hypothesis 1 assumes that well-trained deepfake detection models indicate images based on neither source-relevant nor target-relevant visual concepts, *i.e.* considering them to be artifact-relevant. In this section, we both qualitatively and quantitatively verify the hypothesis.

For the qualitative analysis, we find that artifact-relevant visual concepts barely have intersections with source/target-relevant visual concepts. In Fig. 3, we showed the visual results of $\phi_{v_s}, \phi_{v_t}, \phi_{v_d}$ and the intersections among the main contributed visual concepts for different manipulation algorithms used in FF++ [32]. For the better visualization, we normalized $\phi_{v_s}, \phi_{v_t}, \phi_{v_d}$ all to the unit vector. The backbone of the detection decoder v_d is ResNet-18 [18]. The source and target relevant visual concepts are denoted based on the mask M_τ. For more clarity, in the column of *Intersection*, we only kept the top highest 30% contributed grids. Results show that deepfake detection models mainly consider artifact-relevant concepts as neither source-relevant nor target-relevant.

For the quantitative analysis, we evaluated the proposed metric Q among various DNNs and manipulation algorithms. In Table 1, we calculated the average value of Q among different thresholds τ for a fair comparison. Specifically, τ was set to different values to keep $\{0.60L^2, 0.65L^2, ..., 0.85L^2, 0.9L^2, 0.95L^2\}$ grids on M_τ respectively. $Q > 0$ represents that the learned artifact-relevant visual concepts are more related to source/target-irrelevant visual concepts than the source/target-relevant visual concepts. Results show that various types of DNNs mainly consider artifact-relevant visual concepts as neither source-relevant nor target-relevant. Moreover, such results are not essentially related to the choices on backbones of v_s and v_t, which further verify the generality of the hypothesis. Note that $Q < 0$ for Xception [32] when tested on images manipulated by FaceShifter [23]. It is because that the originally released pre-trained models Xception in [32] was never trained on forged images of FaceShifter [23] before, thus unable to locate the artifact-relevant visual concepts for FaceShifter [23].

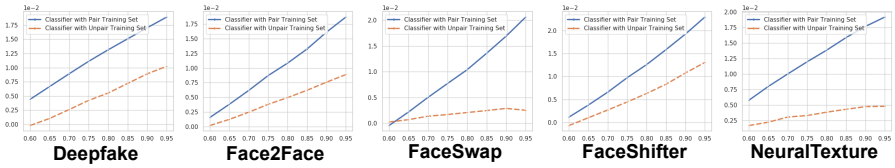

Fig. 4. Verification of hypothesis 2: comparison of the proposed metric Q_τ between models trained on the paired training set and the unpaired train set. The horizontal coordinate represents the percentage of the kept grids in the mask M_τ when setting different thresholds τ. The backbone of the detection encoder is ResNet-18 [18]. Results show that models trained on the paired training set have larger values of Q_τ, showing that FST-Matching helps models to locate artifact-relevant visual concepts.

Table 1. Verification of hypothesis 1: comparison of the proposed metric Q ($\times 10^{-2}$) for different deepfake detection models among various manipulation algorithms. Results show that well-trained deepfake detection models have larger values of Q, which indicates that these models consider source/target-irrelevant visual concepts as artifact-relevant.

Backbone of v_s/v_t	Forgery Methods	Backbone of v_d ($Q(\times 10^{-2})$)				
		ResNet-18	ResNet-34	Efficient-b3	MAT [49]	Xception [32]
ResNet-18 [18]	FaceSwap [21]	2.77	2.88	2.02	2.57	3.10
	Face2Face [40]	2.31	2.63	2.08	2.54	2.59
	FaceShifter [23]	2.45	3.22	2.10	2.42	-0.73
	Deepfake [11]	2.53	2.67	2.30	2.79	2.61
	NeuralTexture [39]	2.30	2.67	2.07	2.51	1.00
Efficient-b3 [38]	FaceSwap [21]	2.85	2.99	2.08	2.49	3.20
	Face2Face [40]	2.19	2.63	2.00	2.49	2.61
	FaceShifter [23]	2.38	3.22	2.07	2.33	-0.67
	Deepfake [11]	2.51	2.71	2.17	2.77	2.64
	NeuralTexture [39]	2.32	2.69	2.05	2.47	1.06

Verification of hypothesis 2. Hypothesis 2 assumes that well-trained deepfake detection models implicitly learned artifact-relevant visual concepts through the FST-Matching in the training set. To verify the hypothesis, we trained two models of the same backbone on the paired training set and the unpaired training set separately. In the paired training set, real images are only the source and target images corresponding to the fake images. In contrast, the real images in the unpaired training set do not match fake images, but are of the same number as the real images in the paired training set. **Both the paired and unpaired training set are downsampled from FF++ [32] dataset containing only 40 identities of images, which is significantly small compared to the initial 1000 identities in the FF++ [32] dataset.** In this section, we conduct extensive experiments to demonstrate that FST-Matching is crucial for learning deepfake detection models.

Firstly, we compared the ACC and video-level AUC on each trained model. As shown in Table 2, models trained on the paired training set achieved similar performance to the baseline models, which are trained on the whole FF++ [32] dataset. Note that the paired training set is significantly smaller than the original FF++ [32] dataset, which demonstrates the importance of FST-Matching in the

Table 2. Verification of hypothesis 2: performance comparison between models trained on the whole FF++ [32] dataset (denoted as the *Baseline*), the paired training set and the unpaired training set. In the paired training set, real images are the corresponding source and target images of fake images *i.e.* satisfying the FST-Matching. Results show that models trained on the paired training set achieve similar performance to the baseline. Note that paired training set is of a significantly smaller size. Such results demonstrate the effectiveness of the FST-Matching.

Models	Forgery Methods	Baseline		Pair		Unpair	
		ACC	*AUC*	*ACC*	*AUC*	*ACC*	*AUC*
ResNet-18 [18]	FaceSwap [21]	98.93	100	97.50	99.91	53.93	75.41
	Face2Face [40]	96.79	99.43	97.14	99.27	64.29	85.74
	FaceShifter [23]	99.29	99.99	97.14	99.82	81.07	93.03
	Deepfake [11]	98.21	100	97.50	99.87	69.64	86.51
	NeuralTexture [39]	90.71	98.89	95.71	98.73	60.00	76.60
Efficient-b3 [38]	FaceSwap [21]	100	100	99.64	100	77.50	87.51
	Face2Face [40]	99.29	99.77	99.29	99.72	81.79	93.36
	FaceShifter [23]	99.29	99.93	99.29	99.96	84.29	96.10
	Deepfake [11]	100	100	100	100	85.36	97.81
	NeuralTexture [39]	99.29	99.85	98.93	99.56	82.86	92.30

Table 3. Verification of hypothesis 3: comparisons between the stability metric δ of different visual concepts. The backbones of the source,target and detection encoders are all ResNet-18 [18]. Results show that learned source and target visual concepts are more consistent to video compression than implicitly learned artifact visual concepts.

Visual Concept	Forgery Methods (δ)				
	FaceSwap	Face2Face	FaceShifter	Deepfake	NeuralTexture
Source	0.73	0.74	0.73	0.74	0.74
Target	0.73	0.76	0.71	0.75	0.76
Artifact (Baseline)	0.17	-0.02	0.14	-0.15	-0.14

training set. In contrast, models trained on the unpaired training set, although of the same size as the paired training set, showed apparently worse results. Such results also show that FST-Matching in the training set is of great value to learning deepfake detection models.

Moreover, we compared the proposed metric Q_τ between each trained model as well. To make a fair comparison, we calculated the value of the metric Q_τ of different τ among all the test images. As shown in Fig 4, models trained on the paired training set have larger values of Q_τ, showing that FST-Matching in the training set effectively helps models to localize source/target-irrelevant visual concepts and consider them as artifact-relevant.

Verification of hypothesis 3. Hypothesis 3 assumes that the implicitly learned artifact visual concepts through the FST-Matching in the raw training set are vulnerable to the video compression. To verify the hypothesis, we tested the raw-trained models on compressed videos and calculated the proposed metric δ_{v_d} among all test images. For the qualitative analysis, as shown in Fig 2, raw-trained models indicate compressed images with significantly different visual concepts compared with the raw images. For the quantitative analysis, in Table

Table 4. Performance comparison on compressed videos with state-of-the-art methods. Our method achieves great performance on compressed videos, especially on c40 videos.

Models	Backbone	C23		C40	
		ACC	*AUC*	*ACC*	*AUC*
Steg.Features [13]	-	70.97	-	55.98	-
LD-CNN [8]	-	78.45	-	58.69	-
Face-x-ray [24]	HRNet	-	87.30	-	61.60
MesoNet [1]	Xception	83.10	-	70.47	-
Xception [32]	Xception	92.39	94.86	80.32	81.76
Xception-ELA [16]	Xception	93.86	94.80	79.63	82.90
Xception-PAFilters [6]	Xception	-	-	87.16	90.20
SPSL [26]	Xception	91.50	95.32	81.57	82.82
MAT-Xception [49]	Xception	96.37	98.97	86.95	87.26
MAT-Efficient [49]	Efficient-b4	**97.60**	**99.29**	88.69	90.40
FST-Matching (ours)	ResNet-18	94.52	98.34	**88.92**	**92.02**
	Xception	94.05	98.27	87.38	90.44
	Efficient-b3	95.95	98.75	87.62	90.89
	Efficient-b4	96.19	98.81	88.69	91.27

3, the calculated $\delta_{v_d} \in [-1, 1]$ is near 0, which also indicates the great change of ϕ_{v_d} under the condition of different compression rate.

Moreover, we also evaluated the stability of the source/target visual concept. Surprisingly, as Fig 2 and Table 3 show, such learned visual concepts show great consistency to the video compression, compared to the implicitly learned artifact visual concepts. Such results motivate us to improve the model performance on compressed videos by devising a model, which explicitly exploits the FST-Matching in the training set.

4.4 FST-Matching Deepfake Detection Model

Performance comparison on compressed videos. In this section, we compared the performance of our model to current state-of-the-art methods. Table 4 shows the performance on compressed videos. Specifically, when aligned with the same backbone of other methods, our model achieved great performance on compressed videos, especially on highly-compressed (*e.g.* c40) videos. Such results also indicate the broad applicability of our method. Meanwhile, note that there still exists a slight performance gap with MAT [49] on c23 in Table 4. Different from our method, MAT [49] designed specific modules to learn the frequency features of images. Such features are widely shown to be effective to enhance the performance of deepfake detection models on compressed videos [14, 22, 26, 28]. To this end, we believe that integrating such features into our model may potentially fill this performance gap. Moreover, since our method is merely the first attempt to exploit our innovative explanation results, we believe that more effective methods could be further inspired based on our study in the future.

Performance comparison on raw videos. In order to have a more comprehensive analysis, we also evaluated our models on raw videos. Results in Table 5 show that our method still performed well on raw images.

Table 5. Evaluation on raw videos.

Models	Backbone	RAW	
		ACC	*AUC*
Face-x-ray [24]	HRNet	-	98.80
MesoNet [1]	Xception	95.23	-
Xception [32]	Xception	**99.26**	99.20
Xception-ELA [16]	Xception	98.57	98.40
MAT-Efficient [49]	Efficient-b4	97.77	99.61
FST-Matching (ours)	ResNet-18	98.14	99.72
	Xception	98.71	99.91
	Efficient-b3	98.93	99.90
	Efficient-b4	99.00	**99.92**

Table 6. Cross-dataset evaluation.

Models	Backbones	Celeb-DF
Xception [32]	Xception	49.03
SPSL [26]	Xception	76.88
MAT [49]	Efficient-b4	68.44
Face-x-ray [24]	HRNet	80.58
FST-Matching (ours)	ResNet-18	86.00
	Xception	88.44
	Efficient-b3	**89.39**
	Efficient-b4	88.13

Table 7. Robustness evaluation to image editing in terms of AUC (%) on FF++.

Method	Saturation	Contrast	Block	Noise	Blur	Pixel	**Avg**
Xception [32]	99.3	98.6	99.7	53.8	60.2	74.2	81.0
Face-x-ray [24]	97.6	88.5	99.1	49.8	63.8	88.6	81.2
LipForensices [17]	**99.9**	99.6	87.4	73.8	96.1	95.6	92.1
FST-Matching (ours)	99.6	**99.9**	**99.9**	**84.8**	**99.2**	**98.7**	**97.0**

Evaluation on the generalization ability. We conduct another experiment to evaluate the generalization ability of our method. To this end, we followed the same cross-dataset experimental setting in SPSL [26]. Results are shown in Table 6, where the metric is AUC (%). Our models trained on FF++ [32] achieved great performance on Celeb-DF [25], regardless of different backbones.

Robustness to image editing operations. We conduct another experiment to evaluate our method when image editing operations are applied to images. To this end, we followed the same robustness experiment setting in LipForensics [17]. Results are shown in Table 7, where the metric is AUC (%). Our method also demonstrated great robustness to listed perturbations.

5 Conclusions

In this paper, we interpret the success of deepfake detection models from the novel perspective of image matching. To this end, three hypotheses are proposed and verified among various DNNs, *i.e.* 1. Deepfake detection models indicate real/fake images based on visual concepts that are neither source-relevant nor target-relevant, that is, considering such visual concepts as artifact-relevant. 2. Besides the supervision of binary labels, deepfake detection models implicitly learn artifact-relevant visual concepts through the FST-Matching in the training set. 3. Implicitly learned artifact visual concepts through the FST-Matching in the raw training set are vulnerable to video compression. Based on the understanding, we further propose the FST-Matching Deepfake Detection Model and achieve great performance on the compressed videos. This research provides an opportunity to explore the essence of artifact representation of images and sheds new light on the task of deepfake detection.

References

1. Afchar, D., Nozick, V., Yamagishi, J., Echizen, I.: Mesonet: a compact facial video forgery detection network. In: 2018 IEEE international workshop on information forensics and security (WIFS). pp. 1–7. IEEE (2018)
2. Ancona, M., Oztireli, C., Gross, M.: Explaining deep neural networks with a polynomial time algorithm for shapley value approximation. In: International Conference on Machine Learning. pp. 272–281. PMLR (2019)
3. Bayar, B., Stamm, M.C.: A deep learning approach to universal image manipulation detection using a new convolutional layer. In: Proceedings of the 4th ACM workshop on information hiding and multimedia security. pp. 5–10 (2016)
4. Castro, J., Gómez, D., Tejada, J.: Polynomial calculation of the shapley value based on sampling. Computers & Operations Research **36**(5), 1726–1730 (2009)
5. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV). pp. 839–847. IEEE (2018)
6. Chen, M., Sedighi, V., Boroumand, M., Fridrich, J.: Jpeg-phase-aware convolutional neural network for steganalysis of jpeg images. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. pp. 75–84 (2017)
7. Cheng, X., Rao, Z., Chen, Y., Zhang, Q.: Explaining knowledge distillation by quantifying the knowledge. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12925–12935 (2020)
8. Cozzolino, D., Poggi, G., Verdoliva, L.: Recasting residual-based local descriptors as convolutional neural networks: an application to image forgery detection. In: Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security. pp. 159–164 (2017)
9. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Canton Ferrer, C.: The deepfake detection challenge dataset. arXiv e-prints pp. arXiv–2006 (2020)
10. Dosovitskiy, A., Brox, T.: Inverting visual representations with convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4829–4837 (2016)
11. FaceSwapDevs: Deepfakes. https://github.com/deepfakes/faceswap (2019)
12. Fong, R., Vedaldi, A.: Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8730–8738 (2018)
13. Fridrich, J., Kodovsky, J.: Rich models for steganalysis of digital images. IEEE Transactions on information Forensics and Security **7**(3), 868–882 (2012)
14. Gu, Q., Chen, S., Yao, T., Chen, Y., Ding, S., Yi, R.: Exploiting fine-grained face forgery clues via progressive enhancement learning. arXiv preprint arXiv:2112.13977 (2021)
15. Guan, C., Wang, X., Zhang, Q., Chen, R., He, D., Xie, X.: Towards a deep and unified understanding of deep neural models in nlp. In: International conference on machine learning. pp. 2454–2463. PMLR (2019)
16. Gunawan, T.S., Hanafiah, S.A.M., Kartiwi, M., Ismail, N., Za'bah, N.F., Nordin, A.N.: Development of photo forensics algorithm by detecting photoshop manipulation using error level analysis. Indonesian Journal of Electrical Engineering and Computer Science **7**(1), 131–137 (2017)
17. Haliassos, A., Vougioukas, K., Petridis, S., Pantic, M.: Lips don't lie: A generalisable and robust approach to face forgery detection. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 5039–5049 (2021)
18. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
19. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 7132–7141 (2018)
20. Hu, Z., Xie, H., Wang, Y., Li, J., Wang, Z., Zhang, Y.: Dynamic inconsistency-aware deepfake video detection. In: IJCAI (2021)
21. Kowalski, M.: FaceSwap. https://github.com/MarekKowalski/FaceSwap (2018)
22. Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y.: Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6458–6467 (2021)
23. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
24. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5001–5010 (2020)
25. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207–3216 (2020)
26. Liu, H., Li, X., Zhou, W., Chen, Y., He, Y., Xue, H., Zhang, W., Yu, N.: Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 772–781 (2021)
27. Lundberg, S.M., Lee, S.I.: A unified approach to interpreting model predictions. Advances in neural information processing systems **30** (2017)
28. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16317–16326 (2021)
29. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5188–5196 (2015)
30. Nguyen, H.H., Yamagishi, J., Echizen, I.: Use of a capsule network to detect fake images and videos. arXiv preprint arXiv:1910.12467 (2019)
31. Rahmouni, N., Nozick, V., Yamagishi, J., Echizen, I.: Distinguishing computer graphics from natural images using convolution neural networks. In: 2017 IEEE Workshop on Information Forensics and Security (WIFS). pp. 1–6. IEEE (2017)
32. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1–11 (2019)
33. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
34. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision. pp. 618–626 (2017)
35. Shapley, L.S.: A value for n-person games, contributions to the theory of games, 2, 307–317 (1953)

36. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034 (2013)
37. Sun, Z., Han, Y., Hua, Z., Ruan, N., Jia, W.: Improving the efficiency and robustness of deepfakes detection through precise geometric features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3609–3618 (2021)
38. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
39. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG) **38**(4), 1–12 (2019)
40. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2face: Real-time face capture and reenactment of rgb videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2387–2395 (2016)
41. Weber, R.J.: Probabilistic values for games. The Shapley Value. Essays in Honor of Lloyd S. Shapley pp. 101–119 (1988)
42. Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., Lipson, H.: Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579 (2015)
43. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)
44. Zhang, D., Zhou, H., Zhang, H., Bao, X., Huo, D., Chen, R., Cheng, X., Wu, M., Zhang, Q.: Building interpretable interaction trees for deep nlp models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 14328–14337 (2021)
45. Zhang, H., Li, S., Ma, Y., Li, M., Xie, Y., Zhang, Q.: Interpreting and boosting dropout from a game-theoretic view. In: International Conference on Learning Representations (2020)
46. Zhang, H., Xie, Y., Zheng, L., Zhang, D., Zhang, Q.: Interpreting multivariate shapley interactions in dnns. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 10877–10886 (2021)
47. Zhang, Q., Cao, R., Shi, F., Wu, Y.N., Zhu, S.C.: Interpreting cnn knowledge via an explanatory graph. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 32 (2018)
48. Zhang, Q., Yang, Y., Ma, H., Wu, Y.N.: Interpreting cnns via decision trees. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
49. Zhao, H., Zhou, W., Chen, D., Wei, T., Zhang, W., Yu, N.: Multi-attentional deepfake detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2185–2194 (2021)
50. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning to recognize patch-wise consistency for deepfake detection. arXiv preprint arXiv:2012.09311 (2020)
51. Zhao, T., Xu, X., Xu, M., Ding, H., Xiong, Y., Xia, W.: Learning self-consistency for deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15023–15033 (2021)
52. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. arXiv preprint arXiv:1412.6856 (2014)
53. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)

54. Zhou, Y., Lim, S.N.: Joint audio-visual deepfake detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14800–14809 (2021)