A Preliminaries: the Shapley value

Originally introduced in game theory [6], the Shapley value was used to distribute the total *award/contribution* obtained by all players to each individual fairly. Specifically, given the set of n input players $N = \{1, 2, ..., n\}$ who participate in the game v, they can obtain the *score* $v(N)$. Here, the game v is formulated as a function to map any participating players to a real number. The *award* obtained by players N is then calculated as $v(N) - v(\emptyset)$, where $v(\emptyset)$ is considered as the baseline *score* when no players participate in the game v. In order to fairly allocate the overall *award*, the Shapley value $\phi(i|N)$ is calculated as the average marginal *award* obtained by player i, when player i joined any potential subset $S \subseteq N \backslash \{i\}$, *i.e.* $v(S \cup \{i\}) - v(S)$. In this way, the Shapley value $\phi(i|N)$ is calculated as follows.

$$\phi_v(i|N) = \sum_{S \subseteq N \backslash \{i\}} \frac{|S|!\,|N-1-S|!}{|N|!}(v(S \cup \{i\}) - v(S)) \tag{1}$$

Moreover, the Shapely value satisfies four properties to ensure its fairness and trustworthiness [10]:

- *Linearity property:* Considering three games u, v and w, where u, v are combined as w. If such games satisfy $w(S) = u(S) + v(S)$, then the Shapley value of each player i in the game w can be combined by the Shapley value of each player i in the game u and the game v, *i.e.* $\phi_w(i|N) = \phi_u(i|N) + \phi_v(i|N)$.
- *Dummy property:* If $v(S \cup \{i\}) - v(S) = 0$ for any subset $S \subseteq N \backslash \{i\}$, then the player i is considered as a dummy player. Its *contribution* is measured as $\phi_v(i|N) = v(\{i\}) - v(\emptyset)$, which indicates that player i participates the game v independently.
- *Symmetry property:* If $v(S \cup \{i\}) = v(S \cup \{j\})$ for any subset $S \subseteq N \backslash \{i, j\}$, then the player i and player j are considered to have the same *contribution*, *i.e.* $\phi_v(i|N) = \phi_v(j|N)$.
- *Efficiency property:* The overall *award/contribution* can be added up by the *award/contribution* of each player i, *i.e.* $\sum_i \phi_v(i|N) = v(N) - v(\emptyset)$.

B More about verification of hypothesis 1

In this section, we provide more results of different backbones to verify the hypothesis. Specifically, following the same setting in Table 1, we used another model, *i.e.* ResNet-34 [2], as the backbone of ϕ_{v_s} and ϕ_{v_t}. Results in Table 8 are consistent with Table 1, indicating that the learned artifact-relevant visual concepts of well-trained deepfake detection models are neither source-relevant nor target-relevant. Such results further support the hypothesis.

Moreover, to further verify the fairness of the proposed metric Q, we evaluated the relationship between the proposed metric Q and the Accuracy (ACC) of deepfake detection models. Specifically, as shown in Fig. 5, values of Q and Accuracy (ACC) of models are positively correlated. Such results show that **when**

deepfake detection models achieve high accuracy, they indicate fake images based on visual concepts, which are neither source-relevant nor target-relevant.

Table 8. More results about verification of hypothesis 1: comparison of the proposed metric Q ($\times 10^{-2}$) for different deepfake detection models among various manipulation algorithms. The backbones of v_s and v_t are all ResNet-34 [2]. Results are consistent with Table 1, which further supports hypothesis 1.

Backbone of v_s/v_t	Forgery Methods	Backbone of v_d ($Q(\times 10^{-2})$)				
		ResNet-18	ResNet-34	Efficient-b3	MAT [11]	Xception [5]
ResNet-34 [2]	FaceSwap [3]	2.67	2.80	2.04	2.53	2.99
	Face2Face [9]	2.07	2.42	1.96	2.51	2.40
	FaceShifter [4]	2.36	3.18	2.14	2.34	-0.68
	Deepfake [1]	2.39	2.57	2.20	2.79	2.49
	NeuralTexture [8]	2.11	2.49	1.93	2.48	0.91

C More about verification of hypothesis 2

In this section, we provide more results of different backbones to verify the hypothesis. Specifically, following the same setting in Table 2, we used ResNet-34 [2] as the backbone and trained two models on the paired training set and unpaired training set respectively. Results in Table 9 are consistent with Table 2, which further support the hypothesis, indicating that the FST-Matching in the training set is of great importance to learn deepfake detection models.

Table 9. More results about verification of hypothesis 2: performance comparison between models trained on the whole FF++ [5] dataset (denoted as the *Baseline*), the paired training set and the unpaired training set. Results are consistent with Table 2, which further demonstrates the effectiveness of the FST-Matching.

Models	Forgery Methods	Baseline		Pair		Unpair	
		ACC	*AUC*	*ACC*	*AUC*	*ACC*	*AUC*
ResNet-34 [2]	FaceSwap [3]	98.93	100	97.14	99.82	68.21	71.76
	Face2Face [9]	98.21	99.26	97.50	99.28	71.07	78.56
	FaceShifter [4]	98.21	99.77	97.50	99.93	79.29	87.21
	Deepfake [1]	98.93	100	99.29	99.99	77.14	82.83
	NeuralTexture [8]	98.21	99.28	96.79	98.26	70.00	77.61

D More about verification of hypothesis 3

In this section, we provide more results of different backbones to verify the hypothesis. Specifically, we followed the same setting in Table 3 and used ResNet-34

[2], EfficientNet-b3 [7] as backbones. Results in Table 10 are consistent with Table 3, indicating that the learned source/target visual concepts are more robust to video compression among different backbones, compared to the implicitly learned artifact visual concepts. Such results further support the hypothesis.

Table 10. More results about verification of hypothesis 3: comparisons between the stability metric δ of different visual concepts. Results are consistent with Table 3 among different backbones, *i.e.* learned source and target visual concepts are more consistent to video compression than implicitly learned artifact visual concepts.

Visual Concept	Backbones	Forgery Methods (δ)				
		FaceSwap	Face2Face	FaceShifter	Deepfake	NeuralTexture
Source (ϕ_{v_s})	ResNet-34 [2]	0.72	0.73	0.72	0.73	0.74
Target (ϕ_{v_t})		0.74	0.76	0.72	0.75	0.76
Artifact (ϕ_{v_d})	ResNet-34 [2]	0.34	-0.02	0.18	0.00	0.04
Source (ϕ_{v_s})	Efficient-b3 [7]	0.65	0.66	0.64	0.66	0.67
Target (ϕ_{v_t})		0.70	0.73	0.63	0.72	0.74
Artifact (ϕ_{v_d})	Efficient-b3 [7]	0.23	0.02	0.13	-0.09	-0.12

E More about the FST-Matching Deepfake Detection Model

E.1 Comparison with the baseline in terms of δ.

In this section, we compared the proposed metric δ between the baseline (*i.e.* the detection encoder v_d) and the FST-Matching Deepfake Detection Model. Results in Table 11 show that compared to the baseline, artifact visual concepts learned by our model are more stable among compressed images. Such results demonstrate the effectiveness of our method.

Table 11. Comparison of the proposed metric δ between the baseline (*i.e.* the detection encoder v_d) and the FST-Matching Deepfake Detection Model. The backbones of the baseline and our model are all ResNet-18 [2]. Results show that compared to the baseline, our model considers more similar visual concepts as artifact-relevant among compressed images.

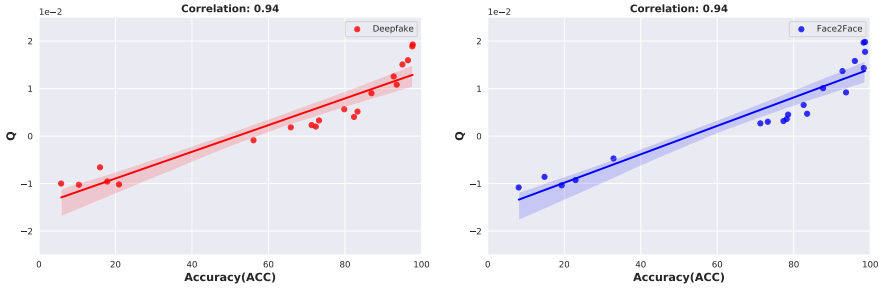
Visual Concept	Forgery Methods (δ)				
	FaceSwap	Face2Face	FaceShifter	Deepfake	NeuralTexture
Artifact (Baseline)	0.17	-0.02	0.14	-0.15	-0.14
Artifact (FST-Matching)	0.54	0.46	0.47	0.45	0.40

E.2 Evaluation of the proposed metric Q.

Besides, we also evaluated the proposed FST-Matching Deepfake Detection Model via the proposed Q to demonstrate its robustness to video compression. All models were trained on the raw dataset and tested on c23 and c40 compressed datasets afterwards. We calculated the proposed metric Q between the baseline model and our model. The backbone of each model is set as ResNet-18 [2]. Results in Table 12 show that our model has a significantly larger value of Q on the c23 and c40 images, indicating the robustness of our model to different compression rates. Note that there exists a performance gap between the baseline and the FST-Matching Deepfake Detection Model on raw images. To this end, compared with the baseline in Table 12, our method is designed to explicitly disentangle the source/target-irrelevant representation from source/target visual concepts on images. Intuitively, such disentangled representation is less enriched than the overall representation of raw images learned by the baseline, causing the performance drop on raw images. However, the disentangled source/target-irrelevant representation is verified to be robust to video compression in the paper, which facilitates our model to achieve great performance on compressed videos.

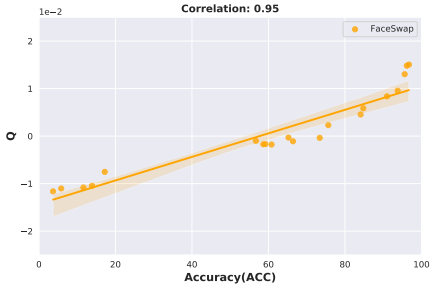
Table 12. Comparison of proposed metric Q ($\times 10^{-2}$) between the baseline and the FST-Matching Deepfake Detection Model. Here Q is averaged among different thresholds τ same as Table 1. Such results show that our model considers similar image regions as artifact-relevant visual concepts among different compressed images.

Forgery Methods	Raw ($Q(\times 10^{-2})$)		C23 ($Q(\times 10^{-2})$)		C40 ($Q(\times 10^{-2})$)	
	Baseline	*FST-Matching*	*Baseline*	*FST-Matching*	*Baseline*	*FST-Matching*
FaceSwap [3]	**2.77**	2.08	0.37	**1.41**	-0.52	**1.15**
Face2Face [9]	**2.31**	1.97	-0.87	**1.30**	-1.35	**0.94**
FaceShifter [4]	**2.45**	2.09	-0.61	**1.05**	-0.67	**0.72**
Deepfake [1]	**2.53**	2.06	-1.64	**1.22**	-1.42	**0.87**
NeuralTexture [8]	**2.30**	1.84	-1.78	**0.71**	-1.44	**0.05**

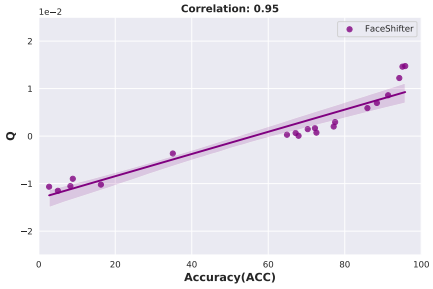


(a) Positive correlation between Q and ACC on Deepfake [1].

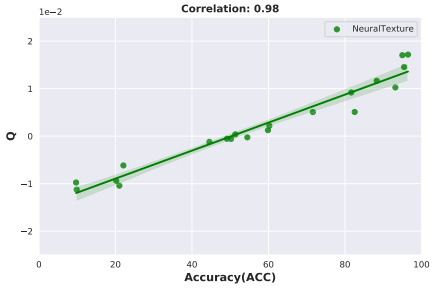
(b) Positive correlation between Q and ACC on Face2Face [9].



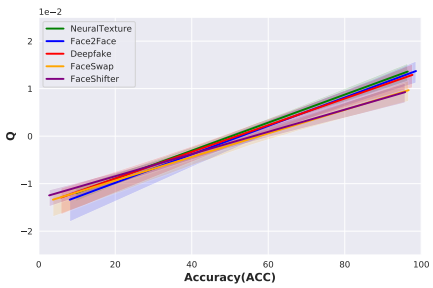
(c) Positive correlation between Q and ACC on FaceSwap [3].



(d) Positive correlation between Q and ACC on FaceShifter [4].



(e) Positive correlation between Q and ACC on NeuralTexture [8].



(f) Positive correlation between Q and ACC on all manipulation algorithms of FF++ [5].

Fig. 5. The positive correlation between the proposed metric Q and the Accuracy (ACC) of deepfake detection models. Different points represent models of different iterations trained on FF++ [5]. The correlation is calculated as the Pearson correlation. The backbones of models are ResNet-18 [2]. Fig. 5(f) shows that the positive correlations between the metric Q and the Accuracy (ACC) are similar among different manipulation algorithms. **Such results show that models with high accuracy consider source/target-irrelevant visual concepts as artifact-relevant.**

References

1. FaceSwapDevs: Deepfakes. https://github.com/deepfakes/faceswap (2019)
2. 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)
3. Kowalski, M.: FaceSwap. https://github.com/MarekKowalski/FaceSwap (2018)
4. 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)
5. 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)
6. Shapley, L.S.: A value for n-person games, contributions to the theory of games, 2, 307–317 (1953)
7. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
8. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics (TOG) **38**(4), 1–12 (2019)
9. 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)
10. Weber, R.J.: Probabilistic values for games. The Shapley Value. Essays in Honor of Lloyd S. Shapley pp. 101–119 (1988)
11. 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)