Face Forgery Detection via Reconstructing Authentic-like Frequency

Zhao-Qian Yuan
National Taiwan University
Taipei, Taiwan
r10921061@ntu.edu.tw

Chia-Mu Yu
National Yang Ming Chiao Tung University
Hsinchu, Taiwan
chiamuyu@gmail.com

I-Cheng Yeh
Yuan Ze University
Taoyuan, Taiwan
ichenyeh@gmail.com

Sy-Yen Kuo
National Taiwan University
Taipei, Taiwan
sykuo@ntu.edu.tw

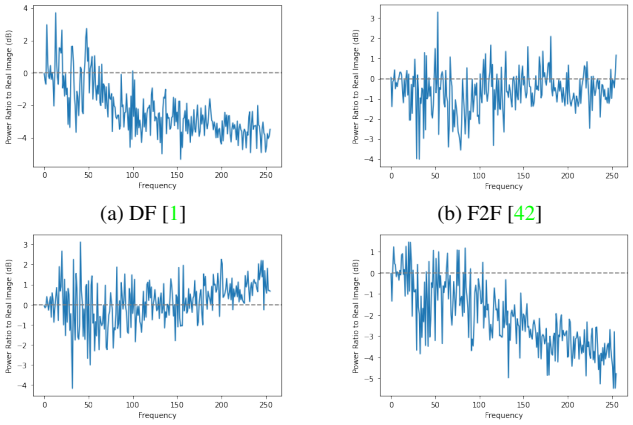
Abstract

Current learning-based face forgery detectors typically discriminate between authentic and synthetic images by capturing artifacts, such as noise or textural discontinuity, caused by manipulation. Unfortunately, these artifact-centric methodologies frequently exhibit constrained generalization capabilities beyond specific databases, making them impractical in real-world scenarios, where novel or unseen artifacts may be present. Recent studies have indicated that facial manipulation often leaves a non-negligible trace in the spectrum. Therefore, we propose a novel forgery detection framework, RALF, that learns the common features of authentic faces and uses the difference between an input image and its reconstructed image to discriminate between real and forged images. RALF initially suppresses suspicious frequencies and subsequently reconstructs the information within the suppressed frequency band to enhance the realism of the reconstruction, thereby improving generalization performance against unknown manipulations. Moreover, we introduce a novel loss function that combines reconstruction loss, metric learning loss, and classification loss to enhance the model's detection capability. We evaluated RALF on several widely used benchmark datasets and demonstrated competitive experimental results compared to state-of-the-art approaches.

1. Introduction

The utilization of deep generative models has significantly proliferated in recent years, thereby democratizing the application of artificial intelligence and enhancing its accessibility to a broader audience. One of the most prominent subdomains within this technological sphere is facial generation and manipulation [1, 2, 28, 41, 42], colloquially known as *deepfake*, which has introduced entertainment and mind-blowing applications. Nevertheless, the nefarious exploitation of these applications, including the dissemination of disinformation and digital identity theft, can engender substantial risks to societal integrity. Given the escalating consciousness regarding privacy and digital security, the development of face forgery detection has become a matter of urgency to prevent its malicious use. Conventional forgery detection methodologies analyze statistical or physical characteristics, such as noise [25] or lighting [21] within the image. While these classical strategies may have exhibited efficacy in nascent stages of deepfake, they currently exhibit vulnerable to sophisticated manipulation techniques.

The advent of Convolutional Neural Networks (CNNs) has catalyzed substantial progress in the realm of facial forgery detection. CNN-based detectors are generally trained on a significant amount of image data, thereby the network can learn to extract features and make inferences based on the corpus of images it has processed. This typically results in remarkable performance during intra-dataset testing. Nevertheless, given the variability of features across different datasets, there is a notable decline in performance when the distribution diverges from that of the training set, leading to a phenomenon known as overfitting. In practical scenarios, it is almost impossible to know in advance which manipulation algorithm employed by the attacker a prior, and it is impractical to retrain the model in response to each novel attacker. Consequently, the model frequently struggles to generalize effectively to the task at hand. As shown in Figure 1, the manipulated image typically diverges from the authentic image within the mid to high-frequency bands.Since studies [14, 45] has indicated that forgery cues within the frequency spectrum of altered images can vary across different generative al-



(a) DF [1] (b) F2F [42]

(c) FS [2] (d) NT [41]

Figure 1. Ratio of the DCT power spectrum of the manipulated data to the pristine data.

gorithms, mere extraction of features from the frequency spectrum may lead to the predicament of over-fitting. From this vantage point, our objective is to discern the ubiquitous features inherent in genuine faces, rather than just merely extracting all frequency-based forgery indicators from the training data.

Overview of RALF. In this paper, we propose a novel forgery detection framework, RALF, that emphasizes the disparity between authentic-like and original images, with the aim of enhancing both performance and robustness shown in Section 4.4. The main idea is to pinpoint the frequency bands typically inhabited by forgeries and subsequently obliterating these suspect frequencies utilizing a learnable mask. We subsequently reconstruct these masked signals to enhance realism within these suspect bands by learning features common to authentic faces. Ideally, the reconstruction of a genuine image should mirror the original, while a forged image should diverge from its realistic reconstruction. Both the original image and its reconstruction are fed into the classifier, guided by the obliterated information, which is deemed as the manipulated region of the image.

Furthermore, to bolster the generalizability of RALF and enhance the delineation between authentic and forged classes, we apply metric learning at the output stages of both the encoder and decoder within the reconstructor. Our methodology has demonstrated encouraging generalization outcomes across multiple datasets.

Contribution. Our contributions are shown as follows:

- We introduce a framework, RALF, using frequency features to enhance forgery detection performance. In particular, RALF obliterates suspect frequencies and reconstructs images, discerning between authentic and forged images.
- Our proposed loss function combines metric learning with traditional losses, further improving discrimination between reconstructions and originals.
- Empirical results highlight RALF's efficacy in detecting various facial forgeries, offering a robust alternative to existing methods.

2. Related Work

Some of the works related to deepfake detection and metric learning are shown in Sections 2.1 and 2.2, respectively.

2.1. Deepfake Detection

We do not intend to present an exhaustive review of deepfake detections; instead, we present only those works closely related to RALF.

Frequency-based Detection. Spatial-frequency analysis has been extensively employed within the realms of image processing and computer vision. Numerous investigations [13, 14, 45] have substantiated that the frequency spectrum of generated images manifests distinct forgery cues within the mid to high-frequency bands. Spatial-frequency transformation techniques transform the image into the frequency domain across disparate components [4, 9] or frequency bands [10]. These transformations facilitate the discernment of minute alterations within each component or frequency band, thereby aiding in the identification of underlying artifacts. Pertinent studies exemplifying these concepts include the following: FAST [50] extracts features across the spatial, temporal, and frequency domains to pinpoint the in-painted regions within the video. SPSL [31] posits that upsampling artifacts, typically manifesting during forgery media generation, exert a pronounced influence on the phase spectra, thereby facilitating the detection of forged images. F^3-Net [35] dissects the image into frequency bands utilizing multiple masks and computes the frequency statistics of each patch to identify statistical alterations in manipulated faces. Luo *et al.* [32] regards noise as a high-frequency feature and uses both color textures and high-frequency noise to capture residual traces. These methodologies predominantly rely on the texture information, either within the spatial or frequency domain, gleaned from the forged samples, consequently constraining their capacity to generalize across a variety of manipulations.

Adversarial-based Detection. Generative Adversarial Networks (GAN) [15] have demonstrated considerable potential in the realm of face forgery detection, attributed to their capacity to generate highly verisimilar images that serve as training data. Face X-ray [29] generates images utilizing a rudimentary blending operation, a common occurrence in deepfake generation, and identifies artifacts induced by the manipulation. SBI [39] can be perceived as a

more challenging variant of Face X-ray, given that it generates images with foreground and background components derived from the identical source image. Chen *et al*. [5] generate realistic facial images with a highly flexible blending step and learn to apprehend all ensuing artifacts. In these methodologies, the discriminator is deployed with the objective of accurately detecting facial forgeries. Nonetheless, the pre-established manipulation processes within the generator, designed to emulate the deepfake algorithm, can impede the discriminator's ability to detect counterfeit images generated with different steps or algorithms. These limitations become apparent when the manipulations deviate from the training set. Additionally, adversarial-based detection necessitates a comprehensive and diverse collection of real, lossless training data, which poses challenges due to their scarcity and the difficulties in acquisition.

2.2. Metric Learning

Metric learning focuses on quantifying the similarity between objects. This technique is commonly employed in self-supervised learning [6] and identification tasks [17,49], where the objective is to determine if objects belong to the same class or cluster. CFL-Net [34] proposes a contrastive learning module that compares the embeddings of real and manipulated pixels to facilitate the identification of the manipulated region. Kumar *et al*. [26] proposes a triplet loss to improve the detection performance in high compression scenario. FDFL method incorporates metric learning [27] to improve the separability of the embedding. By leveraging metric learning, FDFL aims to enhance the ability to distinguish between normal and anomalous facial patterns. Two-branch [33] applies metric learning to separate manipulated faces by pushing them away from the reference center within the feature space. The objective is to create distinct embeddings for manipulated faces, enabling a clearer differentiation from genuine faces.

We enhance the detection capabilities of RALF in differentiating real and fake faces by introducing a novel loss function that integrates metric learning. This approach enables RALF to capture subtle disparities and intricate patterns, improving overall forgery detection efficacy.

3. Proposed Method

3.1. Architecture

We propose RALF, an innovative generalized forgery detection framework that leverages frequency analysis to differentiate between authentic and synthetic faces by reconstructing frequencies akin to those of authentic faces. RALF is composed of four integral components: a frequency destructor, a re-constructor, a discriminator, and a classifier. The frequency destructor is trained to discern the signal constituents that predominantly distinguish counterfeit images from authentic ones, as opposed to merely eliminating high or low-frequency components. Subsequently, the reconstructor then revitalizes the decimated image to augment its authenticity. In the final stage, the classifier, guided by the destructed residual, determines the label based on the divergence between the input image and its reconstructed version. A comprehensive overview of the RALF architecture is illustrated in Figure 2.

Frequency Destructor. Despite the challenge of detecting artifacts in the spatial domain, their noticeable trace in the frequency domain, frequently by attackers, serves as an ideal cue for identifying forged images. The Discrete Cosine Transform (DCT), renowned for its high energy concentration, is extensively utilized in image processing and is thus incorporated into image compression standards. [44,46]. Within RALF, we employ DCT to transform the input image into the frequency domain.

The frequency destructor module, denoted as $\mathcal{D}(\cdot)$, is to pinpoint and obliterate the frequency band encompassing the most informative features for differentiating between authentic and manipulated images. Initially, we transform the input image from the spatial domain to the frequency domain. Then, the frequency spectrum suppressed via a trainable mask, denoted as $\mathcal{M}$. After this destruction process, the residual spectrum is convert back into the spatial domain. The initial process of the input image x is formulated as

$$\hat{x} = \mathcal{D}(x) = \mathcal{F}^{-1}(\mathcal{F}(x) \otimes \mathcal{M}), \tag{1}$$

where $\otimes$ denotes element-wise multiplication and $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ are the forward and inverse DCT operations, respectively. The mask $\mathcal{M} \in [0, 1]$ is a learnable parameter that's fine-tuned to identify and suppress the information in the suspicious frequency band. The destructed image, denoted as $\hat{x}$ and retaining residual frequency components, is utilized in RALF's subsequent stages for input image reconstruction.

Reconstructor. The reconstructor, represented as $\mathcal{R}(\cdot)$, is an encoder-decoder architecture engineered to reconstruct a face $\tilde{x} = \mathcal{R}(\hat{x})$ that mirrors a real face more accurately by assimilating the common features in both the spatial and frequency domains of authentic faces. The encoder extracts features from the destructed input image and compresses it into a lower-dimensional representation, while the decoder reconstructs the image from this compressed representation.

$\mathcal{R}(\cdot)$ is designed to ensure $d(x_r, \tilde{x}_r) \ll d(x_f, \tilde{x}_f)$, where $d(\cdot)$ is a distance function and x_r and x_f are the real and fake inputs, respectively. To further enhance the separation between the input and its reconstruction, we incorporate metric learning at both ends of the encoder and decoder. We incorporate the loss function proposed in the

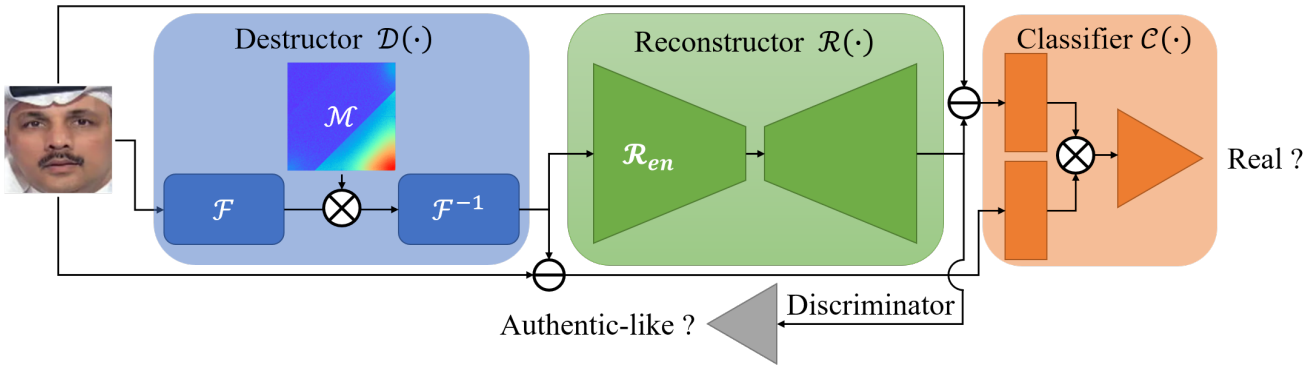


Figure 2. Overview of the proposed RALF framework.

Two-branch [33] into RALF with a specific design.

$$\begin{aligned}\mathcal{L}_{emb} =& \mathbb{E}_{x_r \sim X_r}[\max(\|\mathcal{R}_{en}(\hat{x}_r) - \mathrm{c}\|_2 - \mathrm{r}_r, 0)] \\ &+ \mathbb{E}_{x_f \sim X_f}[\max(\mathrm{r}_f - \|\mathcal{R}_{en}(\hat{x}_f) - \mathrm{c}\|_2, 0)]. \end{aligned} \quad (2)$$

Here, X_r and X_f represent the sets of authentic and fake images in the training set X, respectively.The encoder of our reconstructor is denoted as $\mathcal{R}_{en}(\cdot)$. A reference center in the embedding space is represented as c. The target radius for real and fake faces are denoted as r_r and r_f, respectively, with $\mathrm{r}_r < \mathrm{r}_f$.

More specifically, Two-branch applies $\mathcal{L}_{emb}$ only at the final feature output to isolate manipulated faces. However, in our design, we incorporate $\mathcal{L}_{emb}$ into the embedding space of our reconstructor to initially separate authentic and forged faces. Essentially, $\mathcal{L}_{emb}$ imposes constraints on the input face embeddings, compelling them to aggregate around a reference center, while preventing convergence to a single point and avoid compromising the quality of the reconstruction. The $\mathcal{L}_{emb}$ is illustrated in Figure 3.

Our approach differs from Two-branch in that we dynamically update c during training based on the real faces, rather than precomputing and keeping it fixed during training. This allows us to find an optimal center c that more effectively separates the real and fake faces in the embedding space,

$$\mathrm{c} \leftarrow \mathbb{E}_{x_r \sim X_r}[\mathcal{R}_{en}(\hat{x}_r)]. \quad (3)$$

Spatial reconstruction is a crucial aspect of our design, as it ensures that the generated images resemble real faces as closely as possible. To achieve this, we use a $\mathcal{L}_1$ loss function that allows us to evaluate pixel-level differences between the authentic face $\tilde{x}_r$ and its reconstruction x_r

$$\mathcal{L}_1 = \mathbb{E}_{x_r \sim X_r}[\|\tilde{x}_r - x_r\|_1], \quad (4)$$

which encourages the reconstructor to produce a faithful spatial reconstruction that closely matches the real face.

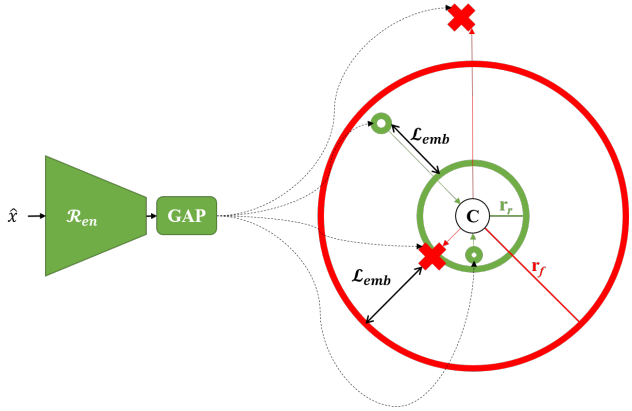


Figure 3. Illustration of $\mathcal{L}_{emb}$. The encoder output is passed through global average pooling (GAP) before being projected into a feature space. In this space, the deepfake image (red cross) contributes to the loss if it is inside the red sphere, while the real sample (green circle) contributes to the loss if it is outside the green sphere.

To guarantee that $\tilde{x}$ encapsulates not only the spatial details but also the frequency information of x, we propose a frequency loss that considers the energy distribution of the DCT coefficients. The DCT coefficients exhibit a highly concentrated energy, with most of the energy concentrated in the low frequency band. However, simply calculating the loss between the reconstructed image and the original one would result in a bias towards the low frequency components, while the high frequency components would contribute minimally to the loss. This is not conducive for forgery detection, as it could result in the neglect of subtle variations in the high-frequency components.

To rectify this problem, the frequency loss must be engineered to assign different weights to the low-frequency and high-frequency components, thereby ensuring an equal contribution from all frequency components to the loss. Thus,

the frequency loss $\mathcal{L}_{freq}$ is defined as

$$\mathcal{L}_{freq} = \mathbb{E}_{x_r \sim X_r}[\|W \otimes (\mathcal{F}(\tilde{x}_r) - \mathcal{F}(x_r))\|_1], \quad (5)$$

where W is a weighting matrix that assigns a smaller weight to the low frequency components and a larger weight to the high frequency components.

Metric learning is employed to prompt the reconstructor to learn and encapsulate the disparity between the original input and its reconstruction. Predominantly, reconstruction methodologies utilize only positive samples, which in our study are real images, to train the network.

Inspired by AECR-Net [48], we aim to fully leverage the information from both authentic and manipulated faces. To accomplish this, we draw the real image and its reconstruction closer to each other and push the fake image and its reconstruction as far apart as possible.

This contrastive learning approach using perceptual loss [20] encourages the reconstructor to produce a more accurate reconstruction that distinguishes between real and fake images. Our proposed objective function for the contrastive loss is formulated as

$$\mathcal{L}_{con} = \mathbb{E}_{x \sim X}\left[\sum_{i=1}^{L} w_i \cdot \frac{\|\phi_i(\tilde{x}_r) - \phi_i(x_r)\|_1}{\|\phi_i(\tilde{x}_f) - \phi_i(x_f)\|_1}\right], \quad (6)$$

where ϕ_i is the i-th layer of the fixed pretrained feature extractor ϕ and w_i is the weight of the corresponding layer. Since we only need to focus on the difference in low-level features, we only use the first L hidden layers of ϕ.

Discriminator. To ensure that the reconstruction of both authentic and forged images have the same distribution as the authentic images, a discriminator is introduced. Unlike the preceding method [5], utilized as a forgery detector in the final stage, the discriminator in RALF is exclusively used to differentiate between real and synthetic images.

The discriminator is trained to maximize the probability of correctly identifying the authentic faces and minimize the likelihood of incorrectly classifying the reconstructed images. The discriminator loss function can be formulated as

$$\begin{aligned}\mathcal{L}_{disc} = &- \mathbb{E}_{x_r \sim X_r}[\log(\mathrm{Disc}(x_r))] \\ &- \mathbb{E}_{x \sim X}[\log(1 - \mathrm{Disc}(\tilde{x}))],\end{aligned} \quad (7)$$

where $\mathrm{Disc}(x)$ represents the degree of authenticity of x. Therefore, the adversarial loss for the reconstructor is obtained as

$$\mathcal{L}_{adv} = -\mathbb{E}_{x \sim X}[\log(\mathrm{Disc}(\tilde{x}))]. \quad (8)$$

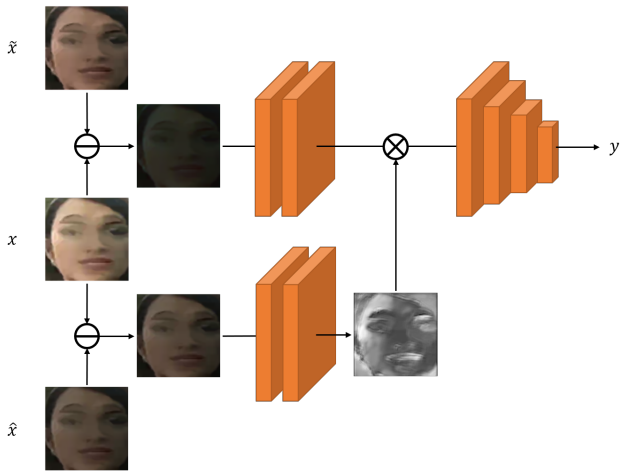


Figure 4. Classifier of RALF. The lower branch is the destruction-guided branch.

$\mathcal{L}_{adv}$ ensures that the distribution of the reconstruction of all images, whether real or fake, is similar to the distribution of the real images that the discriminator cannot tell the difference between them.

We particularly note that the discriminator is solely employed during training and is not utilized during inference. Once the training phase is complete, the discriminator is disregarded and the reconstructor is used to reconstruct images in a standalone manner.

Classifier. The classifier $\mathcal{C}(\cdot)$ is the last component of RALF and determines whether the input image is real or fake by calculating the difference between the input and its reconstruction. To ensure that the classifier concentrates more on the regions affected by the frequency destructor, which are indicative of artifacts, we incorporate a spatial attention module [47] to guide the classifier. The global average pooling layer is replaced by the global max pooling layer to capture subtle differences. The process is illustrated in Figure 4. The classification is formulated as

$$y = \mathcal{C}(|x - \tilde{x}|, |x - \hat{x}|), \quad (9)$$

where y is the prediction of input x. We simply apply binary cross entropy loss to train the classifier:

$$\mathcal{L}_{\mathcal{C}} = -\mathbb{E}_{x_r \sim X_r}[\log(y_r)] - \mathbb{E}_{x_f \sim X_f}[\log(1 - y_f)]. \quad (10)$$

3.2. Loss

We consider multiple losses to improve reconstruction quality in both the spatial and frequency domains, ensuring that the reconstructed faces are realistic and authentic. The total reconstruction loss $\mathcal{L}_{\mathcal{R}}$ is defined as

$$\mathcal{L}_{\mathcal{R}} = \mathcal{L}_1 + \lambda_{freq}\mathcal{L}_{freq} + \lambda_{adv}\mathcal{L}_{adv}. \quad (11)$$

All modules in RALF is jointly optimized by the overall loss $\mathcal{L}$

$$\mathcal{L} = \mathcal{L}_{\mathcal{C}} + \lambda_{\mathcal{R}}\mathcal{L}_{\mathcal{R}} + \alpha \cdot \lambda_{emb}\mathcal{L}_{emb} + \lambda_{con}\mathcal{L}_{con}. \quad (12)$$

The weight α is gradually increased from 0 to 1 during training to increase the penalty of $\mathcal{L}_{emb}$ and to avoid being misled by an incorrect reference center c in the early stages of training. The λ's are the weighting parameters for the corresponding losses.

4. Experiment

4.1. Dataset

We used three datasets, FaceForensics++ (FF++) [37], Celeb-DF [30], and Deepfake Detection Challenge Dataset (DFDC) [12], to evaluate the performance of RALF.

FaceForensics++ (FF++) is the most popular face forgery dataset, containing four types of face manipulation algorithms: DeepFakes [1], Face2Face [42], FaceSwap [2], and NeuralTextures [41], and with three compression levels: raw, c23 (HQ), and c40 (LQ) using the H.264 codec [46]. The dataset contains 1,000 real videos, officially divided into 720 for training, 140 for validation, and 140 for testing, and 4000 manipulated videos. Celeb-DF contains 590 real videos and 5,639 synthetic videos generated with an advanced algorithm. Deepfake Detection Challenge Dataset (DFDC) consists of over 120k videos generated with a diverse set of original videos and several face swapping algorithms such as deepfake autoencoder (DFAE) [12] and StyleGAN [22].

4.2. Evaluation Metrics

Although accuracy (Acc) is the most commonly used metric, it can be misleading in highly imbalanced datasets. Therefore, area under the curve (AUC) and equal error rate (EER) are preferred as they provide a more comprehensive assessment of classifier performance. We note that some of the experimental results of the compared methods are directly excerpted from the original papers and [16].

4.3. Implementation Details

The faces input to RALF are detected and extracted from each video frame using MTCNN [51] and then resized to 256×256. The reconstructor is modified from MIMO-UNet [7], the discriminator follows the architecture of DCGAN [36], and the classifier backbone is Xception [8] pretrained on ImageNet [11]. In Eq. (6), we use pretrained VGG19 [40] as the feature extractor and L is set to 5. The model is jointly trained using the Adam optimizer [23] with a universal learning rate of 0.0001. The λ_{freq} is set to 1, and the other λ's are set to 0.1. Since some datasets are unbalanced between true and false classes, this can potentially mislead our model and degrade its performance. To mitigate this, we use the balance sampler from [24] to balance the data. We only apply random horizontal flipping with a 50% probability for fair comparisons.

Method	FF++ c23 [37]		FF++ c40 [37]	
	Acc (%)	AUC (%)	Acc (%)	AUC (%)
MesoNet [3]	83.10	-	70.47	-
Xception [8]	95.73	-	86.86	-
Face X-ray [29]	-	87.35	-	61.60
Two-branch [33]	96.43	98.70	86.34	86.59
Add-Net [53]	96.78	97.74	87.50	91.01
SPSL [31]	91.50	95.32	81.57	82.82
FDFL [27]	96.69	99.30	89.00	92.40
F^3-Net [35]	97.52	98.10	90.43	93.30
Multi-att [52]	97.60	99.29	88.69	90.40
PEL [16]	**97.63**	**99.32**	**90.52**	**94.28**
RALF	94.71	98.92	83.94	89.06

Table 1. Intra-dataset evaluation

4.4. Experimental Results

4.4.1 Intra-dataset Evaluation

Here, we present the results of the intra-dataset evaluation experiment in Table 1, which shows the competitiveness of RALF. Specifically, RALF achieves an AUC of 98.92% (89.06%), which is only 0.40% (5.22%) lower than the best-performing method when trained on high-quality (low-quality) data. While RALF cannot achieve state-of-the-art (SOTA) performance, we particularly note that intra-dataset evaluation results only present a very limited view of the detection capability due to the inherent assumption that the input samples and training set need to share the same distribution. Thus, we turn to focus on cross-dataset evaluation in Section 4.4.2

4.4.2 Cross-dataset Evaluation

In reality, deepfake detection may encounter unseen and out-of-distribution samples even if it has been trained on datasets in Section 4.1. Thus, the generalizability of deepfake detection becomes critical. To evaluate the generalization ability, in cross-dataset evaluation, we trained RALF on low-quality data from the FF++ dataset, since real-world data is often compressed due to factors such as slow Internet speed or limited storage space. After that, we evaluated RALF on unseen datasets, Celeb-DF [30] and DFDC [12],

Method	Celeb-DF [30]		DFDC [12]	
	AUC (%)	EER (%)	AUC (%)	EER (%)
Xception [8]	60.05	43.19	55.65	46.05
Add-Net [53]	57.83	44.44	51.60	54.77
F^3-Net [35]	67.95	36.76	57.87	44.23
Multi-att [52]	68.64	37.08	63.02	40.98
PEL [16]	69.18	35.69	63.31	40.43
RALF	**71.00**	**34.70**	**63.94**	**39.21**

Table 2. Cross-dataset evaluation.

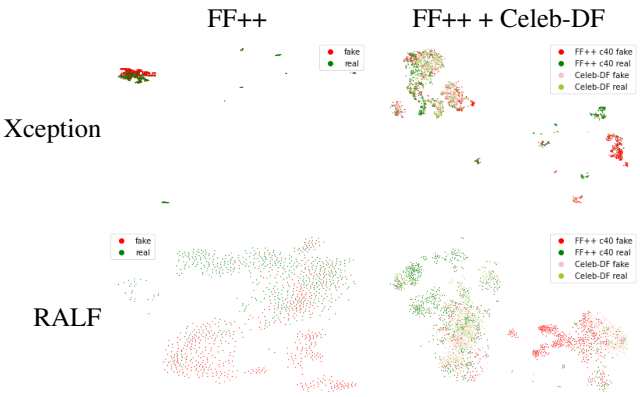


Figure 5. t-SNE visualization of intra-dataset evaluation and cross-dataset evaluation.

which presented a more challenging task. As shown in Table 2, RALF outperforms all prior solutions under this setting. In particular, RALF outperforms the SOTA method PEL [16] by a margin of 1.82% in terms of AUC. To gain further insight, we visualize the feature distributions using t-SNE [43], as shown in Figure 5. In both Xception [8] and RALF, we observe a clear separation between true and false features in the intra-dataset tests. However, there is a clear difference in the embedding space. The feature distribution of Xception is more compact, which may limit its ability to generalize to unknown forgeries, since their projected features can easily lie outside the distribution. In contrast, RALF has a more sparse feature distribution. When unknown forgeries, such as Celeb-DF, are projected into the same feature space as the intra-dataset, Xception struggles to detect the new forgeries, as most of the new forgery samples overlap with the cluster of real features. On the other hand, RALF retains its ability to detect fakes, as some of the new fake samples cluster with the fake intra-dataset. As a whole, Table 2 demonstrates and Figure 5 explains why RALF has a superior generalization capability.

Method	Contrast	Blur	Compression	Pixelation
Xception† [8]	-0.66	-12.76	-3.59	-17.04
F^3-Net† [35]	-2.14	-11.74	-6.72	-18.26
RALF	**-0.40**	**-8.61**	**-2.85**	**-8.37**

Table 3. AUC (%) drop under different distortions. † indicates that the experimental results are derived from our own implementation.

$\mathcal{L}_{freq}$	$\mathcal{L}_{emb}$	$\mathcal{L}_{con}$	AUC (%)
weighted	-	✓	85.43
weighted	✓	-	86.34
-	✓	✓	88.29
unweighted	✓	✓	88.59
weighted	✓	✓	**89.06**

Table 4. Effectiveness of loss component. - and ✓ denote excluded and included, respectively.

4.4.3 Robustness

In the digital world, media often undergoes various distortions during transmission, making the robustness of forgery detectors crucial. We evaluated the robustness of RALF by training it on clean, high-quality images from the FF++ dataset. We then applied second-level distortions, specifically color contrast change, Gaussian blur, JPEG [44] compression, and pixelation, as mentioned in [19]. The impact of these distortions on the performance of different methods is presented in Table 3. Prior methods suffer from a significant drop in performance under these distortions. This is due to their focus on low-level forgery features, which are easily destroyed by such distortions. In contrast, RALF is designed to detect differences between the input and its reconstruction. Since these small distortions are effectively reconstructed, the small differences caused by the corruptions can be ignored. Hence, RALF maintains stable performance even in the presence of distortions.

4.5. Ablation Study

4.5.1 Effectiveness of Loss Component

To demonstrate the effectiveness of the proposed loss component, we conducted variants of RALF with different components excluded. As shown in Table 4, The results highlight the importance of each loss in contributing to the overall effectiveness of RALF. One can also observe from the fourth and fifth rows of Table 4 that by assigning different weights to penalize each frequency component equally, performance can be further improved since the subtle changes in the high frequency band are not ignored.

Backbone	# of Parameters	AUC (%)
ResNet-18 [18]	11.6M	84.30
ResNet-50 [18]	25.5M	85.60
Xception [8]	20.8M	**89.06**

Table 5. RALF of different backbones.

Global Pooling	FF++ [37]	Celeb-DF [30]	DFDC [12]
Average	88.84	69.78	58.24
Max	**89.05**	**71.00**	**63.94**

Table 6. Effectiveness of global pooling layer

4.5.2 Effectiveness of Classifier Backbone

We examined RALF with different backbones on low quality images from FF++. Specifically, we considered ResNet-18 [18] and ResNet-50 due to their exceptional performance in image classification tasks and a similar number of trainable parameters, compared to Xception [8]. As shown in Table 5, it is not surprising that ResNet-18 has a lower AUC score with fewer parameters. Even though ResNet-50 has more learnable parameters, it gains only a small improvement and still performs significantly worse than Xception, which is able to capture fine-grained features, by a substantial margin of 3.46% in terms of AUC. Based on these observations, we chose Xception as our classifier backbone.

4.5.3 Effectiveness of Global Pooling Layer

We investigate the effectiveness of different global pooling layers, specifically global average pooling (GAP) and global maximum pooling (GMP), in RALF. Global pooling is commonly used to reduce the dimensionality of feature maps and improve their representativeness. GAP is preferred in most classification models due to its ability to consider the entire content of the image. In Table 6, both models are trained on the FF++ c40 [37] dataset. We observe that the performance does not vary significantly between the two pooling strategies in the intra-dataset evaluation. However, when we examined the cross-dataset comparison, max pooling showed better performance. This can be attributed to the principle of RALF, which relies on detecting subtle differences between an image and its reconstruction. In such cases, where the difference is often small or inconspicuous, GAP's averaging operation may smooth out these small differences, while GMP emphasizes the most salient and distinctive features but ignores less relevant information, allowing it to effectively capture and highlight these subtle differences.

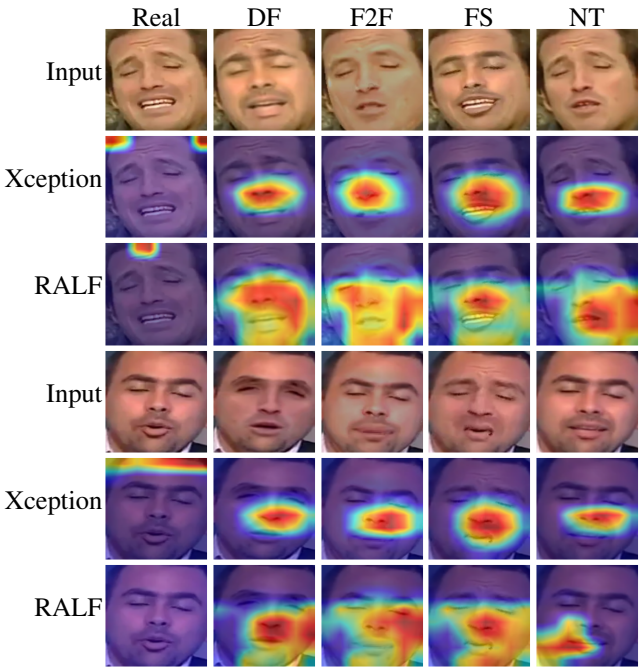


Figure 6. Grad-CAM for the forgery class.

4.5.4 Decision Visualization

We used Grad-CAM [38] to provide insights of how RALF makes the detection. Both Xception [8] and RALF are trained on high-quality data from FF++. The visualized activation maps are shown in Figure 6. The heatmaps highlight areas in the input images that are most relevant to the forgery class. For Xception, the heatmaps consistently show high responses in the center of each image for all types of manipulations. In contrast, RALF shows different response patterns for different types of manipulations; *e.g.* DF [1] leaves noticeable traces in the nose region, while NT [41] typically shows disparities in the mouth region. The different response patterns observed in RALF provide a more explainable and interpretable result compared to Xception. This highlights the fine-grained detection capabilities of RALF.

5. Conclusion

In this paper, we introduce RALF, a novel forgery detection framework that compares input images with authentic-like reconstructions in the frequency domain. By leveraging common features found in real faces, RALF demonstrates improved generalization capabilities, accurately detecting diverse manipulation techniques. Our proposed loss function enhances discriminative power and captures subtle differences, yielding competitive performance and robustness against real-world attacks. The findings contribute to ad-

vancing forgery detection methods, showcasing the potential of frequency-based reconstruction and metric learning for authenticity verification. Future research can explore refinements and extensions of RALF to address emerging forgery techniques and ensure the integrity of digital media content.

References

[1] Deepfakes. https://github.com/deepfakes/faceswap. Accessed: 2023-4-6. 1, 2, 6, 8

[2] Faceswap. https://github.com/MarekKowalski/FaceSwap/. Accessed: 2023-4-6. 1, 2, 6

[3] 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*, pages 1–7. IEEE, 2018. 6

[4] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, Jan 1974. 2

[5] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18710–18719, 2022. 3, 5

[6] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020. 3

[7] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4641–4650, October 2021. 6

[8] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, July 2017. 6, 7, 8

[9] James W Cooley and John W Tukey. An algorithm for the machine calculation of complex fourier series. *Mathematics of computation*, 19(90):297–301, 1965. 2

[10] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on pure and applied mathematics*, 41(7):909–996, 1988. 2

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. 6

[12] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset, 2020. 6, 7, 8

[13] Ricard Durall, Margret Keuper, Franz-Josef Pfreundt, and Janis Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019. 2

[14] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR, 2020. 1, 2

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[16] Qiqi Gu, Shen Chen, Taiping Yao, Yang Chen, Shouhong Ding, and Ran Yi. Exploiting fine-grained face forgery clues via progressive enhancement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 735–743, 2022. 6, 7

[17] Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Is that you? metric learning approaches for face identification. In *2009 IEEE 12th international conference on computer vision*, pages 498–505. IEEE, 2009. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[19] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. DeeperForensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 7

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 5

[21] Micah K Johnson and Hany Farid. Exposing digital forgeries by detecting inconsistencies in lighting. In *Proceedings of the 7th workshop on Multimedia and security*, pages 1–10, 2005. 1

[22] 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. 6

[23] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[24] Sergey Kolesnikov. Catalyst - accelerated deep learning r&d. https://github.com/catalyst-team/catalyst, 2018. 6

[25] Neal Krawetz and Hacker Factor Solutions. A picture's worth. *Hacker Factor Solutions*, 6(2):2, 2007. 1

[26] Akash Kumar, Arnav Bhavsar, and Rajesh Verma. Detecting deepfakes with metric learning. In *2020 8th international workshop on biometrics and forensics*, pages 1–6. IEEE, 2020. 3

[27] Jiaming Li, Hongtao Xie, Jiahong Li, Zhongyuan Wang, and Yongdong Zhang. 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*, pages 6458–6467, June 2021. 3, 6

[28] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 1

[29] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020. 2, 6

[30] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-df: A large-scale challenging dataset for deepfake forensics. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3204–3213, June 2020. 6, 7, 8

[31] Honggu Liu, Xiaodan Li, Wenbo Zhou, Yuefeng Chen, Yuan He, Hui Xue, Weiming Zhang, and Nenghai Yu. Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 772–781, June 2021. 2, 6

[32] Yuchen Luo, Yong Zhang, Junchi Yan, and Wei Liu. Generalizing face forgery detection with high-frequency features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16317–16326, 2021. 2

[33] Iacopo Masi, Aditya Killekar, Royston Marian Mascarenhas, Shenoy Pratik Gurudatt, and Wael AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 667–684. Springer, 2020. 3, 4, 6

[34] Fahim Faisal Niloy, Kishor Kumar Bhaumik, and Simon S. Woo. Cfl-net: Image forgery localization using contrastive learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4642–4651, January 2023. 3

[35] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII*, pages 86–103. Springer, 2020. 2, 6, 7

[36] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015. 6

[37] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *International Conference on Computer Vision*, 2019. 6, 8

[38] 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. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 8

[39] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022. 2

[40] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[41] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *Acm Transactions on Graphics*, 38(4):1–12, 2019. 1, 2, 6, 8

[42] 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*, June 2016. 1, 2, 6

[43] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 7

[44] G.K. Wallace. The jpeg still picture compression standard. *IEEE Transactions on Consumer Electronics*, 38(1):xviii–xxxiv, Feb 1992. 3, 7

[45] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8695–8704, 2020. 1, 2

[46] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003. 3, 6

[47] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European Conference on Computer Vision*, September 2018. 5

[48] Haiyan Wu, Yanyun Qu, Shaohui Lin, Jian Zhou, Ruizhi Qiao, Zhizhong Zhang, Yuan Xie, and Lizhuang Ma. Contrastive learning for compact single image dehazing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10551–10560, 2021. 5

[49] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Deep metric learning for person re-identification. In *2014 22nd international conference on pattern recognition*, pages 34–39. IEEE, 2014. 3

[50] Bingyao Yu, Wanhua Li, Xiu Li, Jiwen Lu, and Jie Zhou. Frequency-aware spatiotemporal transformers for video inpainting detection. In *2021 IEEE/CVF International Conference on Computer Vision*, pages 8168–8177, 2021. 2

[51] 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, 2016. 6

[52] Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2185–2194, 2021. 6, 7

[53] Bojia Zi, Minghao Chang, Jingjing Chen, Xingjun Ma, and Yu-Gang Jiang. Wilddeepfake: A challenging real-world dataset for deepfake detection. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20,

page 2382–2390, New York, NY, USA, 2020. Association for Computing Machinery. 6, 7