Benchmarking Joint Face Spoofing and Forgery Detection With Visual and Physiological Cues

Zitong Yu, Senior Member, IEEE, Rizhao Cai, Zhi Li, Wenhan Yang, Member, IEEE, Jingang Shi, and Alex C. Kot, Life Fellow, IEEE

Abstract—Face anti-spoofing (FAS) and face forgery detection play vital roles in securing face biometric systems from presentation attacks (PAs) and vicious digital manipulation (e.g., deepfakes). Despite satisfactory performance upon large-scale data and powerful deep models, recent advances in face spoofing and forgery detection approaches usually focus on 1) unimodal visual appearance or physiological (i.e., remote photoplethysmography (rPPG)) cues; and 2) separated feature representation for FAS or face forgery detection. On one side, unimodal appearance and rPPG features are respectively vulnerable to high-fidelity face 3D mask and video replay attacks, inspiring us to design reliable multi-modal fusion mechanisms for generalized FAS. On the other side, there are rich common features across FAS and face forgery detection tasks (e.g., periodic rPPG rhythms and vanilla appearance for bonafides), providing solid evidence to design a joint FAS and face forgery detection system in a multi-task learning fashion. In this paper, we establish the first joint face spoofing and forgery detection benchmark using both visual appearance and physiological rPPG cues. To enhance the rPPG periodicity discrimination, we design a two-branch physiological network using both facial spatio-temporal rPPG signal map and its continuous wavelet transformed counterpart as inputs. To mitigate the modality bias and improve the fusion efficacy, we conduct a weighted batch and layer normalization for both appearance and rPPG features before multi-modal fusion. We also investigate prevalent deep models, feature fusion strategies and multi-task learning configurations for joint face spoofing and forgery detection. We find that the generalization capacities of both unimodal (appearance or rPPG) and multi-modal (appearance+rPPG) models can be obviously improved via joint training on these two tasks. We hope this new benchmark will facilitate the future research of both FAS and deepfake detection communities.

Index Terms—Face anti-spoofing, face forgery detection, deepfake, rPPG, fusion, joint training, multi-task learning.

I. INTRODUCTION

Due to its convenience and remarkable accuracy, face recognition [1] technology has been applied in a few interactive intelligent applications such as checking-in and mobile payment. However, existing face recognition systems are vulnerable to presentation attacks (PAs) ranging from print, replay, makeup, 3D-mask, etc. Therefore, both academia and industry have paid extensive attention to developing face anti-spoofing (FAS) [2] technology for securing the face recognition system, which has been widely used in daily applications such as border/gate access control and mobile face payment. In the early stage, according to the bonafide/spoof discrepancy on facial appearance [3] (e.g., high visual quality for the bonafide vs. quality distortion with artifacts for the spoof) and physiological dynamics [4] (e.g., periodic heartbeat rhythm for the bonafide vs. noisy remote photoplethysmography (rPPG) [5] signals for the spoof), plenty of traditional handcrafted visual texture/quality features [6], [7] and physiological liveness cues [4], [8], [9], [10] have been proposed for FAS. Subsequently, thanks to the emergence of large-scale public FAS datasets [11], [12], [13], [14] with rich attack types and recorded sensors, more and more deep learning based methods are proposed for capturing discriminative and generalized visual appearance [15], [16], [17], [18] and physiological rPPG [19], [20] features.

Besides the real-world face spoofing attacks, with unrestricted access to the rapid proliferation of digital face images on social media platforms, launching forgery attacks against face recognition systems has become even more accessible. Given the growing dissemination of ‘fake news’ and ‘deepfakes’ [21], both research community and social media platforms are pushing towards generalizable detection techniques against sophisticated forgery attacks (including entire face synthesis, attribute manipulation, and face swap), which has been widely used in daily media applications such as filtering fake news and frauds. Early works are based on traditional digital forensics features [22], [23] while later ones rely heavily on deep learning based image and video classification (real vs. fake) [24], [25] using the visual appearance deepfake inputs. However, deepfake images and videos are continuously evolving and become more and more visually realistic due to the more mature and powerful deep image generation methods [26]. To tackle the detection issue under weak and unperceived appearance artifacts, physiological rPPG based methods built on traditional time-frequency statistics [27], [28] and deep rhythm features [29], [30], [31] are proposed.
Despite excellent achievements in visual appearance or physiological rPPG based unimodal face attack detection, multi-modal models on both appearance and rPPG cues attract less attention, which might be due to the lack of mature fusion strategies [4], [8]. As for the FAS and face forgery detection tasks, facial appearance features cannot generalize well under unseen scenarios of artifact disappearing while hidden rPPG clues are less reliable when large head motions are included and heavy compression is employed. In this case, detection based on either visual appearance or physiological rPPG cues might be sub-optimal. To the best of our knowledge, there is only one work [8] to explore the feasibility of combining appearance and rPPG predictions via score-level fusion for generalized FAS. However, it is still unknown: 1) how multi-modal models perform on the face forgery detection task; and 2) whether sophisticated feature-level fusions help boost the performance.

In terms of the relationship between FAS and face forgery detection tasks [32], bonafides reasonably share the similar clean appearance and periodic rPPG cues while spoofs/deepfakes have different kinds of artifacts and noisy rhythms. The visual and physiological features from these two tasks are subtle and easily influenced by the undesired domain shifts (e.g., face quality and identity attributes). Despite task-aware differences between physical materials and digital artifacts, features from these two tasks can be treated as out-of-distribution along bonafides, which might be synergistic. The tight correlations in both visual [33] and physiological modalities inspire us to explore whether multi-task learning benefits the common and intrinsic feature representation in both unimodal (appearance or rPPG) and multi-modal (appearance+rPPG) settings against unseen domain shifts. Meanwhile, there are practical scenarios where both face spoofing and deepfake detection tasks are needed. For example, digital facial manipulation might be used before generating a realistic physical ID card for harmful fraud, where face spoofs and forgeries should be detected simultaneously.

Motivated by the above discussions, we propose a joint face spoofing and forgery detection benchmark using both visual appearance and physiological rPPG cues. The frameworks of face attack detection based on the inputs of facial appearance, rPPG, and appearance+rPPG are shown in the first (red), second (blue), and third (purple) columns of Fig. 1, respectively. From the perspective of separate/joint attack detection, tasks shown in the first and second rows of Fig. 1 represent the separate FAS and face forgery detection, respectively. And the joint face spoofing and forgery detection task is illustrated in the last row of Fig. 1, where the bonafide, spoofing, and deepfake samples are all included for training. In this paper, we generally define the bonafide as live samples from FAS task and genuine ones from face forgery detection task, while the spoofing and deepfakes denote physical presentation attacks in FAS task and digital manipulation attacks in face forgery detection task, respectively.

Compared with well-studied appearance-based FAS and face forgery detection methods, the ones based on rPPG or appearance+rPPG [8] cues might be still immature. On one hand, the periodicity discrimination of the time-domain rPPG signals is limited due to external interference (e.g., video quality and head movement). In other words, under serious interference, the rPPG curves from the bonafide might still hold noisy temporal shapes like those from attacks. On the other hand, due to the modality bias (performance and distribution) between the appearance and rPPG features, the direct fusion is unstable and might has negative effects. To tackle the first issue, an extra network branch with continuous wavelet transformed time-frequency representation is introduced, which is able to directly provide useful

![Fig. 1. Visualization of the modality (visual appearance and physiological rPPG) and task (face spoofing and forgery detection) matrix. Columns from left to right: 1) appearance models with RGB face inputs; 2) rPPG models with facial rPPG signals inputs; and 3) appearance+rPPG models with both RGB face and facial rPPG inputs. Rows from top to bottom: 1) separate face spoofing detection with bonafide/spoof for training; 2) separate face forgery detection with bonafide/deepfakes for training; and 3) joint face spoofing and forgery detection with bonafide/spoof/deepfakes for training. Best views in color.](image-url)
dynamic frequency variation cues to enhance the periodicity discrimination. Furthermore, to alleviate the modality bias and improve the fusion efficacy, we propose to adopt a weighted batch and layer normalization for separate modality features refinement before fusion.

In terms of multi-task learning for joint face spoofing and forgery detection, we conduct elaborate studies on prevalent deep models and learning configurations, and find some valuable conclusions: 1) joint face spoofing and forgery detection can benefit the generalization capacities of most prevalent deep models and both modalities (i.e., appearance and rPPG) on cross-dataset testings for both tasks; 2) task sampling strategy influences the multi-task learning a lot, and the simplest random sampling between two tasks surprisingly works best; 3) joint training with both bonafide and attack samples performs the best while with only extra attack samples from the other task might degrade the performance; and 4) backbone sharing in low and mid levels benefits the multi-task learning.

To sum up, our contributions include:

- We establish the first joint face spoofing and forgery detection benchmark where both visual appearance and physiological rPPG cues are fully exploited.
- We design a two-branch physiological network using both facial spatio-temporal rPPG signal map and its continuous wavelet transformation counterpart as inputs to enhance the rPPG periodicity discrimination.
- We conduct a weighted batch and layer normalization for both appearance and rPPG features before multi-modal fusion to mitigate the modality bias and improve the fusion efficacy.
- We investigate prevalent deep models, fusion strategies and multi-task learning configurations for joint face spoofing and forgery detection, and find that the generalization capacities of both unimodal (appearance or rPPG) and multi-modal (appearance+rPPG) models can be obviously improved via joint training on both tasks.

In the rest of the paper, Section II provides the related work. Section III formulates the joint FAS and face forgery detection. Then, it introduces the two-branch physiological framework with CWT map and the fusion strategy with weighted batch and layer normalizations. Section IV provides the details of the benchmark establishment. Section V provides rigorous ablation studies and evaluates the performance of mainstream models on the proposed benchmark. Finally, a conclusion with future directions is given in Section VI.

II. RELATED WORK

In this section, we first introduce the recent progress in FAS based on visual appearance and physiological rPPG cues. Then, previous appearance and rPPG based face forgery detection approaches are investigated. Finally, related multi-task learning methods for face attack detection are presented.

A. Face Anti-Spoofing

Visual Appearance Based FAS: Traditional face anti-spoofing (FAS) methods usually extract hand-crafted texture descriptors, e.g., LBP [34] and HOG [6], from the facial images to capture the spoofing patterns. Some deep learning based methods treat the visual appearance based FAS problem as a binary classification task [35], and utilize binary cross-entropy loss to optimize the model. Motivated by physical discrepancy between bonafide and spoof faces, dense pixel-wise supervisions [36], e.g., pseudo depth map [15], [19], [37], [38], reflection map [11], [16], texture map [39] and binary map [12], [40] are introduced for learning intrinsic features. FAS can be also formulated as domain adaptation or generalization problems in [41], [42], [43]. Densely tangled learning [39], [44], adversarial learning [45], and meta learning [46], [47] are usually adopted to improve the model’s generalization capacity on unseen scenarios. Besides, several anomaly detection [48], [49], zero-shot [12], [50] and continuous learning [51] approaches are proposed for appearance based unknown PA detection.

Physiological rPPG Based FAS: As facial-video based remote rPPG measurement techniques achieve great developments [52], [53], many researchers utilize rPPG to detect face spoofing attacks. Li et al. [4] first consider the difference of pulse between bonafide faces and printed faces for spoofing attack detection. Several works [9], [10], [54], [55] compare facial rPPG signals and background noises to decide whether the faces contain liveness clues or not. Moreover, combining with deep learning methods, Yu et al. [20] propose an rPPG transformer framework (TransRPPG) to extract global periodicity for FAS. Liu et al. [12] utilize the generated pseudo rPPG frequency distributions as auxiliary supervision signals to benefit discriminative dynamic clues representation. Instead of using only the rPPG modality, Lin et al. [8] explore the feasibility of combing appearance and rPPG predictions via score-level fusion for generalized FAS.

B. Face Forgery Detection

Visual Appearance Based Face Forgery Detection: Many efforts have been made to improve the performance of face forgery detection [21], [24], [56], [57]. Early works [58], [59] use generic deep models (e.g., Xception [60]) to extract features from cropped face images and perform binary classification, which is weak in capturing fine-grained artifacts between bonafide and forged images. In view of that forged faces become more visually realistic, more reliable forgery patterns such as noise statistics, local textures, and frequency information are proposed. For example, Zhou et al. [61] design a two-stream neural network to combine visual appearance and local noise patterns for forgery detection. Zhao et al. [62] propose a multi-attentional face forgery detector to efficiently aggregate the texture and high-level semantic features from multiple local regions. Several other works [25], [63], [64] also consider the frequency details into account and propose frequency-aware models to distinguish bonafide faces from forged faces.

Physiological rPPG Based Face Forgery Detection: Similar to the rPPG based FAS methods, rPPG based face forgery detection [28], [65] mainly relies on the judgment about whether the heart rhythm patterns contain solid live/periodic patterns since the heart rhythms extracted from forged face videos are

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
usually diminished by deepfake methods, and are with less periodic cues compared with those from bonafide faces. Ciftci et al. [27] extract rich expert-designed statistical rPPG features from both the time domain and power spectra domain to describe spatial coherence and temporal consistency for deepfake detection. Subsequently, several transformed rPPG maps such as PPG cell [31], motion-magnified spatio-temporal map [29], and Gaussian spatial decomposition map [30] are proposed, which are cascaded with deep CNN or transformer models for bonafide/forgery binary classification.

C. Multi-Task Learning for Face Attack Detection

Multi-task learning aims at developing models that can address multiple tasks via sharing partial parameters and computation [66]. Numerous works have observed that although multi-task models are more versatile, their accuracies are easily lower than single-task models, especially when simultaneously performing unrelated tasks [67], [68]. Moreover, jointly training models to simultaneously perform multiple tasks has typically required careful calibration of the individual tasks, to ensure that none of the task-specific losses dominates another. Methods to mitigate this issue include gradient-normalization [69] and adaptive loss weights [70].

In the field of FAS, few works [16], [71] are proposed to mine task-shared intrinsic features (i.e., material property of generic objects and spoof instruments) via multi-task learning between visual material recognition and FAS tasks. However, there are still huge task-aware biases (e.g., appearance and physiological cues) between generic objects and human faces. As for face forgery detection, some works [72], [73] consider multi-task learning on both detection, forged region segmentation, and blending type/ratio prediction to improve generalization under cross-dataset evaluation. In terms of unified digital and physical face attack detection, one recent work [33] proposes to adaptively learn joint appearance representations for coherent attacks, while uncorrelated attack types are learned separately. All previous works focus on the visual appearance modality but ignore the rPPG based multi-task learning. To fill in the blank, we establish the first joint face spoofing and forgery detection benchmark using both visual appearance and physiological rPPG cues to explore whether more common and intrinsic features can be learned on such two highly-correlated tasks.

III. METHODOLOGY

In this section, we first introduce the formulation of joint face spoofing and forgery detection in Section III-A. To enhance the rPPG periodicity discrimination, we then propose a two-branch physiological network in Section III-B. Finally, to improve the fusion efficacy between appearance and rPPG features, we present a simple yet effective weighted batch and layer normalization strategy in Section III-C.

Preliminary: Followed by most previous works, we treat face spoofing and forgery detection tasks as a binary (bonafide/attack) classification problem to train models with binary cross-entropy (BCE) loss [59], [74] or binary pixel-wise supervision [36], [40]. Given an extracted face $X_{App}$ or rPPG signal map $X_{rPPG}$ input, deep appearance features $F_{App}$ or rPPG features $F_{rPPG}$ can be represented via forwarding the corresponding unimodal models $\Phi_{App}$ and $\Phi_{rPPG}$, respectively. Then the cascaded classification heads make the binary predictions $Y_{Pre}$, which are supervised by the BCE loss

$$L_{BCE} = -(Y_{GT} \log(Y_{Pre}) + (1 - Y_{GT}) \log(1 - Y_{Pre})).$$

where $Y_{GT}$ is the ground truth ($Y_{GT} = 0$ for attack and $Y_{GT} = 1$ for bonafide). Similarly, for the models [40] supervised with binary pixel-wise supervision, the BCE loss is accumulated on all spatial dimensions.

A. Joint Face Spoofing and Forgery Detection

In general, separate models (i.e., $\Phi_{Spoof}$ and $\Phi_{Forgery}$) are trained for face spoofing and forgery detection. Such separate training for each task usually neglects the high correlation between these two tasks. On one hand, bonafides reasonably share the similar clean appearance and periodic rPPG cues. Learning upon the merged bonafide samples from both two tasks is able to provide more valid samples to alleviate overfitting for bonafide representation. On the other hand, spoofs and deepfakes have different kinds of artifacts and noisy rhythms, which enhances the diversity of the attack data. Thus, learning from the merged diverse attack samples might improve the models’ generalization. Instead of training within each standalone task, we propose to train a unified model $\Phi_{Joint}$ in a multi-task learning fashion.

Joint Training Architecture: Here we consider the simplest setting (i.e., all backbones and classification heads are shared across two tasks) first. In this case, there are no extra learnable task-specific parameters thus it is as efficient as single-task models for inference. Specifically, the shared heads with 2 classes (bonafide/attack) and 3 classes (bonafide/spoof/forgery) are shown in Fig. 2(a) and (b), respectively. Compared with the former one supervised via the BCE loss and making a binary prediction, the latter one further distinguishes the attack class into two more fine-grained task-specific classes (i.e., spoof and forgery), and is supervised via the 3-class cross-entropy loss. Besides the shared head setting, there is another favorite setting

![Fig. 2. Different head settings for multi-task learning with (a) a shared single head for binary classification; (b) a shared single head for 3-class classification; and (c) separate heads for binary classification. Task #1 and #2 indicate face spoofing and forgery detection, respectively.](image-url)
to cascade task-specific heads behind the shared model (see Fig. 2(c)) to flexibly project the shared embedding features into a separate decision space. Thus, the classification heads are the only task-specific parameters. The study about the head setting can be found in Table V.

To increase model representation capacity (with N layers) when jointly training on two tasks simultaneously, we can optionally include $N_{\text{shared}} > 0$ task-shared layers in front while cascading $N_{\text{specific}} = N - N_{\text{shared}}$ task-specific layers. In this case, lower-level detailed features (e.g., edge, contour, color and quality) are shared across two tasks to alleviate overfitting while task-specific higher-level semantic features (e.g., moire pattern and spatial inconsistency) are learned separately to enhance the discrimination ability. We study the impacts of $N_{\text{shared}}$ for both appearance and rPPG models in Fig. 5(a)–(d).

**Joint Task Sampling Strategy:** We optimize the unified model parameters $\theta_{\text{joint}}$ simultaneously across two tasks with stochastic gradient descent (SGD). As a result, there are different sampling strategies for task-related training batch construction. During joint training, for each SGD step, we sample a mini-batch from the merged two-task data $D_{\text{Spoof}} \cup D_{\text{Forgery}}$, where $D_{\text{Spoof}}$ and $D_{\text{Forgery}}$ respectively denote the whole training data of FAS task and face forgery task, then evaluate a gradient for updating $\theta_{\text{joint}}$. An important consideration is the proportion of per-task data in each mini-batch and task sampling order. We describe several task sampling schedules below. $U_{\text{Spoof}}$ and $U_{\text{Forgery}}$ are denoted as the total SGD steps within each epoch for single-task face spoofing and forgery detection, respectively.

- **Random.** In this schedule, each batch is randomly sampled from the merged two-task data $D_{\text{Spoof}} \cup D_{\text{Forgery}}$, thus the proportion of per-task data is stochastic.
- **Simultaneous.** Here we sample the minibatch always encompassing data from both tasks simultaneously. Specifically, a half of the minibatch is assembled with data from $D_{\text{Spoof}}$ while the other half is from $D_{\text{Forgery}}$.
- **Alternating.** This deterministic schedule alternates between $D_{\text{Spoof}}$ and $D_{\text{Forgery}}$ in a fixed, repeating order. Concretely, we perform a single SGD step for each task in sequence before repeating the same order.
- **Task-by-task.** In this schedule, data from first $U_{\text{Spoof}}$ SGD steps are used for training in FAS task. Then data from the next $U_{\text{Forgery}}$ steps with face forgery detection task are used for training and vice versa.

Random sampling is adopted as the default minibatch construction strategy due to its simplicity and satisfactory performance. Detailed studies can be found in Table VI.

**B. Two-Branch Physiological Network**

The amplitude of optical absorption variation of face skin is very small, thus it is important to design a good representation to highlight the physiological information in face videos for robust face spoofing and forgery detection. To this end, we utilize both the multi-scale spatial-temporal map (MSTmap) \[\text{[75]}\] and the continuous wavelet transformation (CWT) based time-frequency map as two-branch inputs for the physiological network. The overall framework is illustrated in Fig. 3.

**MSTmap and WaveletMap:** The MSTmap considers both the local and the global physiological information to represent the facial skin color variations due to heartbeats. For the $t$-th frame of a face video, we first get a set of $K$ informative regions of face $R_i = \{R_{1i}, R_{2i}, \ldots, R_{Ki}\}$. Then, we calculate the average pixel values of each RGB color channel for all the non-empty subsets of $R_i$, which are $2K - 1$ combinations of the elements in $R_i$. Therefore, given a facial video with $T$ frames, an MSTmap with the size of $(2K - 1) \times T \times 3$ is obtained, where the three dimensions mean the numbers of total regions, frames, and color channels.

Besides the representation in the time domain, we also exploit the wavelet representation of rPPG signals to form a time-frequency view, which contains straightforward and rich periodicity patterns. Specifically, the CWT of the global rPPG signal $S(t)$ (i.e., averaged from all $K$ facial regions) corresponds to a time-frequency representation computed from a prototype function, i.e., mother wavelet, which can be formulated as

$$CWT^{\psi}_{S}(\tau, d) = \int_{-\infty}^{\infty} S(t)\psi_{\tau,d}(t)dt,$$

$$\psi_{\tau,d}(t) = \frac{1}{\sqrt{|d|}} \psi(t - \tau),$$

(2)

where $\psi_{\tau,d}$ corresponds to the mother wavelet dilated by $d$ and translated by $\tau$. Unlike the Fourier transform, the wavelet transform can detect abrupt changes in frequency using a family of wavelets $\psi_{\tau,d}$ computed from the mother wavelet $\psi$. We use the Matlab implementation ‘cwtfilterbank’ with Morse wavelet and 48 voices per octave.

**Two-Branch Network:** Unlike the previous works \[\text{[20], [29]}\], only considering the time-domain input, the proposed two-branch network adopts the extracted facial MSTmap and WaveletMap as inputs. As shown in Fig. 3, the MSTmap and WaveletMap first forward through the same but parameter-unshared encoders to obtain the time-domain and time-frequency-domain deep features, respectively. Then these two kinds of features are concatenated to form the generalized rPPG features. Finally, a binary classification head supervised by the BCE loss $L_{BCE}^{\text{rPPG}}$ is cascaded for prediction. To benefit the training stability, we also add auxiliary binary classification heads for each separate branch, which are supervised by the BCE loss.
losses $\mathcal{L}_{BCE}^{\text{MSTmap}}$ and $\mathcal{L}_{BCE}^{\text{Wavelet}}$, respectively. The overall loss is denoted as $\mathcal{L}_{\text{overall}}^{\text{PPG}} = \mathcal{L}_{BCE}^{\text{PPG}} + \alpha \cdot (\mathcal{L}_{BCE}^{\text{MSTmap}} + \mathcal{L}_{BCE}^{\text{Wavelet}})$, where hyperparameter $\alpha = 0.5$. By leveraging both the fine-grained time-domain color change rhythms and the semantic time-frequency distribution, the two-branched physiological network is able to capture discriminative and generalized hidden periodicity cues.

C. Weighted Normalization for Appearance and rPPG Fusion

As for the appearance+rPPG multi-modal models, the naive framework is to fuse the appearance features $F_{\text{App}}$ and rPPG features $F_{\text{rPPG}}$ first, and then a prediction head with the BCE loss $\mathcal{L}_{BCE}^{\text{Fuse}}$ is cascaded behind the fused multi-modal features $F_{\text{Fuse}}$. The overall loss for the multi-modal model supervision can be formulated as $\mathcal{L}_{\text{overall}}^{\text{Fuse}} = \mathcal{L}_{BCE}^{\text{Fuse}} + \beta \cdot (\mathcal{L}_{BCE}^{\text{App}} + \mathcal{L}_{BCE}^{\text{rPPG}})$, where hyperparameter $\beta = 0.5$.

Due to the modality discrepancy between visual appearance and physiological rPPG, their corresponding hidden features ($F_{\text{App}}$ and $F_{\text{rPPG}}$) usually differ significantly, making the fusion challenging (e.g., the direct feature summation drops the performance compared with unimodal models). One straightforward solution is to concatenate these two features in the channel domain first [76], and then aggregate the multi-modal heterogeneous features with a lightweight fusion operator (e.g., linear fully connection layer). The direct concatenation fusion can be formulated as

$$F_{\text{Fuse}} = \text{ReLU}(\text{Linear}(\text{Concat}(F_{\text{App}}, F_{\text{rPPG}}))).$$

However, feature misalignment among modalities leads to inefficient feature concatenation fusion. To alleviate this issue, we propose to utilize weighted batch normalization (BN) [77] and layer normalization (LN) [78] for each independent modality branch first to map $F_{\text{App}}$ and $F_{\text{rPPG}}$ into the aligned subspace. Then the refined features ($F_{\text{Norm}}^{\text{App}}$ and $F_{\text{Norm}}^{\text{rPPG}}$) are concatenated and aggregated. The weighted normalization fusion can be formulated as

$$F_{\text{Fuse}} = \text{ReLU}(\text{Linear}(\text{Concat}(F_{\text{Norm}}^{\text{App}}, F_{\text{Norm}}^{\text{rPPG}}))),$$

where hyperparameter $\theta \in [0, 1]$ tradeoffs the contribution between BN and LN, which is studied in Fig. 5(e) and (f). As the embedding features $F_{\text{App}}$ and $F_{\text{rPPG}}$ are already projected among all spatial dimensions (e.g., average pooling in ResNet while global class token in ViT), the feature statistics (i.e., mean and variance) from BN and LN are only calculated for each individual channel across all batch samples and each individual sample across all channels, respectively. The merits of the weighted normalization fusion are two-fold: 1) towards efficient fusion based on the refined modality features via leveraging informative statistics from two views (batch and channel); and 2) enhancing the stability and benefiting the convergence for multi-modal model training.

To verify the efficacy of the weighted batch and layer normalization for appearance and rPPG feature fusion, two other fusion strategies (i.e., squeeze-and-excitation (SE) fusion [79], [80] and cross-attention fusion [81]) are compared (see Table IV). Please note that in this paper, we focus on feature-level fusion strategies but there are also e.g., input-level fusion and decision-level fusion for multi-modal scenarios, which might be explored in future work.

IV. JOINT FACE SPOOFING AND FORGERY DETECTION BENCHMARK

In this section, we introduce the joint face spoofing and forgery detection benchmark in terms of datasets, protocols, and evaluation metrics. A statistical description of the datasets is shown in Table I.

Datasets: In consideration of the efficacy of the long-term physiological rPPG cues [9], [20], we select nine datasets (i.e., SIW [19], 3DMAD [82], HKBU-MarsV2 (HKBU) [9], MSU-MFSD (MSU) [83], 3DMask [84] and ROSE-YouTu (ROSE) [85] for FAS while FaceForensics++ (FF++) [59], DFDC [86] and Celeb_DFv2 [87] for face forgery detection) containing long videos (mostly >5 seconds) for the joint face spoofing and forgery detection benchmark, which is suitable for investigations of both visual appearance and rPPG modalities. The whole MSU, 3DMask, ROSE, DFDC, and Celeb_DFv2 datasets are used for testing directly while the original train/test splits in [19], [59] are strictly followed for SIW and FF++ datasets. In consideration of no original splits for 3DMAD and HKBU datasets, we simply split the samples from the first 11 and 6 subjects for training while the remaining ones for testing. Representative visual samples as well as their rPPG maps are illustrated in Fig. 4.

Protocols: To fairly evaluate the efficacy of the benchmark baselines, we design the protocols under two perspectives: 1) task concurrence; and 2) generalization to domain shifts. As for the former one, separate training on each single task vs. joint training within both two tasks is conducted for verifying the concurrence of face spoofing and forgery detection tasks as well as the effectiveness of multi-task learning. In terms of the latter one, both intra- and cross-domain\(^1\) tests are performed to evaluate the generalization capacities of models.

\(^1\)cross-domain indicate 1) spoof/forgery unrelated external changes (e.g., lighting, sensor noise, and video quality) but actually influence the appearance quality; and 2) the unknown spoofing/forgery attacks usually mean the novel attack types with intrinsic physical properties and digital traces which have not occurred in the training phase.

\[\text{Table I}\]

DATASETS AND PROPERTIES USED IN THE JOINT FACE SPOOFING AND FORGERY DETECTION BENCHMARK

| Dataset          | Train Videos (Bona./Att.) | Test Videos (Bona./Att.) | # Sub. | Attack Types |
|------------------|--------------------------|--------------------------|--------|--------------|
| FAS              | 500/2950                 | 500/2950                 | 10     | FAS          |
| FF+ [9]          | 720/2880                 | 140/560                  | 5      | Face2Face, FaceSwap, Deepfakes, Neural/Textual |
| DFDC [86]        | 131/4133                 | 131/4133                 | 5      | Deepfakes, FaceSwap |
| Celeb_DFv2 [87]  | 597/5639                 | 597/5639                 | 5      | Deepfakes    |

Sub., Bona., and Att. are short for subjects, bona fide and attack, respectively. ‘DFDC’ is short for DFDC preview.
Separate training: This protocol is implemented via training two separate models on datasets \{SiW, 3DMAD, HKBU\} and \{FF++\} for face spoofing and forgery detection tasks, respectively. The trained models are tested on both intra-domain (i.e., \{SiW, 3DMAD, HKBU\} and \{FF++\}) and cross-domain (i.e., \{MSU, 3DMask, ROSE\} and \{DFDC, CelebDFv2\}) scenarios for two tasks, respectively.

Joint training: Instead of training separate models, the joint training protocol is to train a unified model for two tasks via leveraging all data in \{SiW, 3DMAD, HKBU, FF++\}. Similarly, the trained unified model is evaluated on both intra-domain and cross-domain scenarios for two tasks. Note that the cross-domain testing for both separate and joint protocols are more challenging as the testing set covers more unseen datasets and more complex unknown attacks, which are correlated to real-world scenarios.

Evaluation Metrics: For all experiments, Area Under Curve (AUC) and Equal Error Rate (EER) [88] are utilized for performance evaluation. Besides, the evaluations under intra- and cross-domain scenarios among multiple (merged) datasets are investigated by using the metric of single-side True Positive Rate (TPR)@False Positive Rate (FPR) [89], which is more suitable for realistic spectacles.

V. EXPERIMENTS
A. Implementation Details

For each video clip, eight face images are uniformly cropped using MTCNN [91] face detector. We follow [75] to extract the MSTmap with \(K = 6\) and \(T = 300\). In terms of \(T\), the first 300 frames are used when the length of video \(T'\) is larger than 300. If \(T'\) is smaller than 300 frames, we complement the last \(300 - T'\) frames. \(\theta = 0.8\) is utilized for the weighted normalization fusion. The experiments are implemented with Pytorch on one NVIDIA V100 GPU. We use the SGD optimizer with the learning rate (lr) 0.001 and batch size 64 at the training stage. We train models with 30 epochs while lr decays with 0.1 in the 20th epoch. For the appearance-based unimodal experiments, we adopt ResNet50 [92] and ViT-Base [93] as our backbone models for ablation studies, and compare them with mainstream FAS (CDCN++ [15] and DeepPixel [40] with pixel-wise BCE loss) and face forgery detection methods (MesoNet [94], Xception [59], and MultiAtten [62] with BCE loss). In terms of the rPPG-based unimodal experiments, ResNet18 [75] and ViT-Base [93] are used as our backbone models, and the classical FAS method TransRPPG [20] is compared. Finally, ResNet50+ResNet18 and ViT+ViT are used as backbones for the appearance+rPPG multi-modal experiments. All ResNet and ViT initialize with ImageNet1K pretrained weights. All inputs are resized to the defaulted input sizes according to different models. For simplicity, all the joint training results are based on the fully shared backbone with 1 head 2 classes setting except the joint ViT models with name suffix ‘(shared 8)’ only sharing the first 8 transformer blocks.

B. Comparison Results on the Proposed Benchmark

Results of Appearance-Based Unimodal Models: As shown in Table II, it is obvious to find that compared with separate training, most of the appearance models with joint training obtain performance gains especially for challenging cross-testing scenarios for both two tasks. Specifically, remarkable improvements can be found when joint training and cross-testing on 3DMask dataset (see the column ‘3DMask’) on all models, i.e., 1.33%, 11.61%, 21.43%, 13.88%, 9.78%, 20.69%, and 27.14% AUC improvement for MesoNet, Xception, MultiAtten, CDCN++, DeepPixel, ResNet50, and ViT, respectively. From the perspective of testing on merged datasets (see the last block in Table II), we can find that joint training usually makes little difference (e.g., \(\leq 4\%\) TPR@FPR=10% for DeepPixel, ResNet50 and ViT) on intra-tests for both FAS and face forgery detection tasks but gives remarkable improvements on cross-tests (e.g., 9.77%, 5.32% and 7.49% TPR@FPR=10% for MultiAtten, ResNet50 and ViT on face spoofing detection while 6.78%, 5.96% and 6.74% TPR@FPR=10% for DeepPixel, ResNet50 and ViT on face forgery detection). It indicates that learning with more paired bonafide/attack data from the extra task alleviates the
overfitting issue and improves models’ domain generalization capabilities.

In terms of impacts of the neural architectures, we can find that generic models with stronger representation capacities (e.g., ResNet50/ViT vs. lightweight MesoNet) usually benefit more from joint training. The best cross-testing performance can be achieved by ViT(shared 8) model. Specifically, joint training cannot improve the generalization capacity (not significant or even worse) for task-specialized models (MesoNet/MultiAtten in face forgery detection while CDCN++ in face spoofing detection). The reasons behind might be two-fold: 1) compared with generic models (e.g., ResNet/ViT) already verified their effectiveness in both tasks, task-specialized models might induce performance task-aware biases during joint training; and 2) simple joint training strategy does not consider the adaptive adjustment for such task-aware biases thus might damage the generalization capacity for task-specialized architectures. Thus, it is careful to choose suitable models and design reasonable learning strategies for joint training.

Results of rPPG-Based Unimodal Models: As can be seen in Table III, similar to appearance-based unimodal models, joint training usually enhances rPPG models’ generalization under cross-domain scenarios but makes limited improvements in intra-testings. Specifically, obvious improvements can be found when joint training and cross-testing on MSU, 3DMask and DFDC datasets on all models. For instance, when separate training only on three face spoofing detection datasets recorded in lab environments, the rPPG models generalize poorly on 3DMask due to the serious unseen illumination and head movement interference. This issue is alleviated via introducing large-scale in-the-wild rPPG samples from FF++ for joint training. However, the overall performance of rPPG-based models still has a large gap with appearance-based models. The best rPPG model ‘ViT_STW(shared 8)’ with joint training only achieves <40%, <85% and <30% TPR@FPR = 10% for cross-domain FAS, intra-testing and cross-domain face forgery detection, respectively.

Results of Appearance+rPPG Based Multi-Modal Models: Table IV shows the results of appearance+rPPG models with separate training and joint training. Specifically, ResNet50/ViT-Base and ResNet18/ViT-Base are used as the backbones of appearance and rPPG modalities, respectively. On one hand, compared with separate training, joint training on all models with different fusion strategies (except ViT_CA) improve the cross-testing performance on both tasks. Specifically, with

| Metrics        | Sep. / Joint Method | Face Spoofing Detection | Face Forgery Detection |
|---------------|---------------------|--------------------------|------------------------|
|               |                     | Intra-testing            | Cross-testing          | Intra-testing            | Cross-testing          |
|               |                     | SW 3DMAD 3DSU            | MSU 3DMask 3Dセックス | TP++ FPDC Celeb-3DFace  |

|               | MesoNet             | 76.75 ± 3.81            | 78.75 ± 7.1            | 73.03 ± 0.77             | 74.92 ± 0.77             | 31.1 ± 11.92             |
|---------------|---------------------|--------------------------|------------------------|--------------------------|--------------------------|--------------------------|
|               | Xception            | 97.13 ± 3.73            | 97.65 ± 3.33            | 97.65 ± 3.33             | 97.65 ± 3.33             | 30.0 ± 13.6             |
|               |                     | 93.05 ± 3.99            | 93.05 ± 3.99            | 93.05 ± 3.99             | 93.05 ± 3.99             | 30.0 ± 13.6             |
|               | CDCN++              | 98.69 ± 9.95            | 98.69 ± 9.95            | 98.69 ± 9.95             | 98.69 ± 9.95             | 30.0 ± 13.6             |
|               | DeepPixel           | 97.07 ± 3.97            | 97.07 ± 3.97            | 97.07 ± 3.97             | 97.07 ± 3.97             | 30.0 ± 13.6             |
|               |                     | 97.07 ± 3.97            | 97.07 ± 3.97            | 97.07 ± 3.97             | 97.07 ± 3.97             | 30.0 ± 13.6             |
|               | ViT                 | 98.57 ± 9.77            | 98.57 ± 9.77            | 98.57 ± 9.77             | 98.57 ± 9.77             | 30.0 ± 13.6             |
|               |                     | 98.57 ± 9.77            | 98.57 ± 9.77            | 98.57 ± 9.77             | 98.57 ± 9.77             | 30.0 ± 13.6             |
|               | R-PSP              | 72.11 ± 3.25            | 72.11 ± 3.25            | 72.11 ± 3.25             | 72.11 ± 3.25             | 30.0 ± 13.6             |
|               | MesoNet             | 79.36 ± 3.33            | 79.36 ± 3.33            | 79.36 ± 3.33             | 79.36 ± 3.33             | 30.0 ± 13.6             |
|               | Xception            | 94.85 ± 7.51            | 94.85 ± 7.51            | 94.85 ± 7.51             | 94.85 ± 7.51             | 30.0 ± 13.6             |
|               | MultiAtten          | 99.31 ± 9.47            | 99.31 ± 9.47            | 99.31 ± 9.47             | 99.31 ± 9.47             | 30.0 ± 13.6             |
|               | CDCN++              | 98.86 ± 9.87            | 98.86 ± 9.87            | 98.86 ± 9.87             | 98.86 ± 9.87             | 30.0 ± 13.6             |
|               | DeepPixel           | 98.86 ± 9.87            | 98.86 ± 9.87            | 98.86 ± 9.87             | 98.86 ± 9.87             | 30.0 ± 13.6             |
|               | ViT                 | 99.19 ± 9.31            | 99.19 ± 9.31            | 99.19 ± 9.31             | 99.19 ± 9.31             | 30.0 ± 13.6             |
|               |                     | 99.19 ± 9.31            | 99.19 ± 9.31            | 99.19 ± 9.31             | 99.19 ± 9.31             | 30.0 ± 13.6             |
|               | R-PSP              | 99.94 ± 9.31            | 99.94 ± 9.31            | 99.94 ± 9.31             | 99.94 ± 9.31             | 30.0 ± 13.6             |

*Sep.* is short for separate training. The higher AUC(%) *" and TPR(%) *" while the lower FPR(%) the better performance. Best cross-testing results are marked in bold.
TABLE III  
RESULTS OF rPPG-BASED UNIMODAL MODELS ON THE JOINT FACE SPOOFING AND FORGERY DETECTION BENCHMARK

| Metrics | Sep. / joint | Method | Face Spoofing Detection | Face Forgery Detection |
|---------|--------------|--------|-------------------------|------------------------|
|         |              |        | Intra-testing | Cross-testing | Intra-testing | Cross-testing |
| SW      |             |        | SW | JDAU | TRBuv | RGG | Mask | JDAU | TRBuv | RGG | Mask |
| Sep.    | HazeNetPTL_ST | 84.41 | 92.72 | 89.58 | 94.74 | 87.92 | 89.44 | 85.97 | 93.79 | 89.54 | 93.97 | 89.79 | 94.33 |
| AUC(%)  | ResNet18_ST  | 93.61 | 97.42 | 94.06 | 97.60 | 93.76 | 97.76 | 93.97 | 98.44 | 94.97 | 98.77 | 94.97 | 99.73 |
|         | ResNet18_W  | 94.26 | 96.63 | 94.91 | 97.54 | 94.75 | 97.85 | 94.97 | 98.77 | 95.24 | 98.95 | 95.53 | 99.73 |
|         | ResNet18_STW | 95.37 | 97.28 | 94.84 | 97.36 | 95.28 | 97.58 | 95.37 | 99.24 | 95.97 | 99.52 | 96.23 | 99.97 |
|         | ViT_ST      | 96.53 | 98.66 | 95.77 | 98.68 | 96.55 | 98.78 | 96.76 | 99.34 | 97.05 | 99.65 | 97.31 |
|         | ViT_W       | 97.47 | 98.76 | 96.46 | 98.78 | 97.46 | 98.78 | 97.64 | 99.44 | 98.12 | 99.75 | 98.44 |
| Joint   | HazeNetPTL_ST | 83.3 | 93.11 | 80.97 | 85.40 | 83.8 | 93.88 | 84.17 | 97.66 | 83.8 | 97.66 |
|         | ResNet18_ST  | 92.66 | 97.32 | 93.62 | 97.57 | 92.87 | 97.78 | 93.17 | 98.57 | 93.47 | 98.87 | 93.57 | 99.17 |
|         | ResNet18_W  | 93.42 | 98.43 | 94.37 | 98.45 | 93.67 | 98.67 | 93.87 | 99.35 | 94.17 | 99.65 | 94.47 | 99.97 |
|         | ResNet18_STW | 94.23 | 95.93 | 94.94 | 96.73 | 94.75 | 96.89 | 94.97 | 99.24 | 95.97 | 99.52 | 96.23 | 99.97 |
|         | ViT_ST      | 96.56 | 97.83 | 95.43 | 98.66 | 96.21 | 98.75 | 96.24 | 99.25 | 96.56 | 99.52 | 96.86 | 99.97 |
|         | ViT_W       | 97.03 | 98.98 | 96.85 | 98.78 | 97.71 | 98.84 | 97.95 | 99.63 | 98.24 | 99.95 | 98.55 | 99.97 |
|         | ViT_STW     | 97.26 | 98.17 | 96.94 | 98.78 | 97.59 | 98.85 | 97.86 | 99.67 | 98.34 | 99.95 | 98.67 | 99.97 |
|         | ViT_STW(share 8) | 97.53 | 98.28 | 96.94 | 98.82 | 97.82 | 98.90 | 97.95 | 99.69 | 98.44 | 99.96 | 98.74 | 99.97 |
| Joint   | HazeNetPTL_ST | 84.8 | 93.43 | 82.79 | 86.85 | 84.95 | 94.10 | 85.17 | 98.35 | 84.95 | 98.35 | 85.47 | 98.97 |
|         | ResNet18_ST  | 14.09 | 13.33 | 13.69 | 14.67 | 14.33 | 14.67 | 14.77 | 15.67 | 15.33 | 15.97 | 15.77 | 16.33 |
|         | ResNet18_W  | 12.28 | 8.42  | 12.39 | 14.21 | 13.82 | 14.21 | 14.42 | 15.32 | 14.67 | 15.32 | 15.67 | 16.33 |
|         | ResNet18_STW | 11.53 | 6.67  | 13.17 | 14.33 | 16.06 | 15.33 | 15.42 | 16.21 | 15.67 | 16.67 | 16.33 | 17.33 |
|         | ViT_ST      | 9.18  | 13.33 | 9.52  | 14.22 | 13.82 | 14.22 | 14.42 | 15.32 | 14.67 | 15.32 | 15.67 | 16.33 |
|         | ViT_W       | 7.23  | 12.35 | 9.86  | 14.87 | 14.17 | 14.87 | 15.07 | 15.97 | 15.32 | 16.67 | 16.33 | 17.33 |
|         | ViT_STW     | 6.05  | 10.82 | 9.52  | 14.52 | 14.17 | 14.52 | 14.73 | 15.67 | 14.67 | 16.67 | 16.33 | 17.33 |
|         | ViT_STW(share 8) | 5.97 | 10.82 | 9.52  | 14.52 | 14.17 | 14.52 | 14.73 | 15.67 | 14.67 | 16.67 | 16.33 | 17.33 |

C. Ablation Studies

Efficacy of the Two-Branch Physiological Network: The results of rPPG-based models w/ and w/o two-branch inputs are shown in Table III. We can find that in most cases, models ‘ViT_W’ using only Wavelet map perform slightly better than models ‘ViT_ST’ using MSTmap. Compared with models (ResNet18_ST and ViT_ST) with uni-branchn MSTmap input and models (ResNet18_W and ViT_W) with uni-branchn wavelet input, models (ResNet18_STW and ViT_STW) with two-branchn MSTmap and Waveletmap inputs achieve better performance for both intra- and cross-testings on two tasks, indicating that semantic time-frequency representation benefits the rPPG models learning more discriminative and generalized periodic features while the effective mining of the spoof/forgery clues from both spatio-temporal and wavelet domains via two-branchn fusion. Besides, the two-branchn physiological networks are compatible with separate and joint training, consistently improving the performance compared with only MSTmap inputs.

Efficacy of the Weighted Normalization Fusion: According to Equation 4, \( \theta \) controls the contribution of the LN and BN for normalized feature fusion. As illustrated in Fig. 5(e)–(f), when \( \theta = 0.4 \) and 0.8, the multi-modal models (ResNet/ViT architectures) assembled with weighted normalization fusion generalize well for cross-domain face spoofing and forgery detection. Based on the optimal setting with \( \theta = 0.8 \), we can find from Table IV that the weighted normalization fusion consistently performs well when plugged-and-played in different architectures (e.g., ResNet and ViT). In contrast, it can be seen
from the cross-testing results of ‘ResNet_CA’ and ‘ViT_CA’ that the CA fusion works well for ResNet but severely decreases the performance of ViT with joint training. We can also find from the results of ‘ResNet_SE_Concat’ and ‘ViT_SE_Concat’ that SE fusion is suitable for ViT but perform poorly for ResNet with joint training.

It can be seen from Fig. 5(e) and (f) that TPRs are greatly fluctuating for the hyper-parameter $\theta$. The reasons behind might be two-fold: 1) the training of the proposed multimodal method simultaneously updates the network parameters of each branch and the fusion parameters in weighted normalization module, which makes the optimization not well stable thus easily obtaining fluctuating results using different hyper-parameter $\theta$; and 2) using fixed hyper-parameter $\theta$ for all epochs might be not optimal as the importance of statistics features from BN and LN of each modality might be different during different training stages. The fluctuation is more obvious in the case of ViT architecture than ResNet, which might be due to the uses of all LN layers within ViT but all BN layers within ResNet.

**Impacts of Joint Training Architectures:** Table V shows the results of different configurations of the classification heads for joint training. We can see that when the whole backbones (ResNet and ViT) are shared, training with 1 head 2 classes (bonafide/attack) and 2 heads 2 classes settings usually perform well for the ResNet and ViT, respectively. It is interesting to find that compared with separate training, there are some negative effects with the 1 head 3 classes (bonafide, spoof, forgery) setting, indicating that the fine-grained class supervision signal might confuse the joint training.

Besides the head settings, it is necessary to explore the impacts of shared architectures with joint training. Fig. 5(a), (b), (c), and (d) illustrates the results of joint training architectures with different shared blocks using 2 heads 2 classes setting. It can be seen from Fig. 5(a) and (b) that sharing all the residual blocks in ResNet has the best generalization capacity on both tasks and modalities. In terms of ViT architecture in Fig. 5(c) and (d), we find that sharing the first 8 transformer blocks for joint training holds the best generalization capacity, indicating the low- and
mid-level feature sharing is important for efficient joint training. In conclusion, the best joint training configuration for ResNet is to share the whole backbone with 1 head 2 classes setting while sharing the first 8 transformer blocks with 2 heads 2 classes setting for ViT.

**Impacts of Sampling Strategies for Joint Training:**

The results of four sampling strategies across two tasks mentioned in Section III-A are shown in Table III. It is surprising to find that utilizing the simplest sampling strategy, i.e., randomly sampling minibatch data from two tasks, performs the best in terms of modality (appearance and rPPG) and architecture (ResNet and ViT). Compared with the ‘Random’ and ‘Simultaneous’ sampling strategies, the ‘Alternating’ and ‘Task-by-task’ easily result in unstable predictions for intra- and cross-testings, which might be caused by the frequent and contradictory knowledge transfer across tasks.

**Impact of Bonafide/Attacks and Tasks for Joint Training:**

In previous joint training settings, bonafides and attacks from both tasks are fully exploited to learn more generalized representation. Here we explore two biased situations (i.e., only bonafides or attacks from other tasks for joint training) to see how extra bonafides or attacks contribute to the joint training. As shown in Table VII, it is interesting to find that compared with separate training, joint training with only bonafides or attacks is still able to benefit cross-testing face spoofing and forgery detection performance for both modalities except the rPPG-based ‘ViT_STW’ model. Note that bonafides from FAS and face forgery detection tasks are easier to collect in real-world cases, which is more practical for joint training.

Besides the tasks related to face attack detection, it is necessary to see whether the other face analysis tasks benefit the joint training for face spoofing and forgery detection. Here we joint train models with face expression recognition (7 classes of basic expressions) and emotional expression recognition (9 classes of basic expressions).
emotions), gender recognition (male/female/unsure), and race recognition (Caucasian/African-American/Asian) tasks with shared 1 head multiple classes setting. The large-scale facial expression database RAF-DB [90] is used. Table VIII shows the results of joint training face attack detection and face expression/gender/race recognition tasks. Despite acceptable accuracy in the extra expression/gender/race recognition tasks (>80%), there are limited gains for the desired face spoofing and forgery detection. Due to the low correlation among these tasks, training with more face data even drops the performance (e.g., joint training with gender recognition for ViT) compared with separate training. In other words, the benefits of joint training across face spoofing and forgery detection tasks are not only from the extra data volume but also from their high task correlations.

D. Feature Visualization and Discussion

Distributions of the appearance feature $F_{\text{App}}$ from ViT w/ and w/o joint training are shown in Fig. 6 via t-SNE [97]. It is clear that the features with joint training present more well-clustered behavior than that with separate training under cross-domain scenarios. Specifically, the feature distributions of the high-fidelity face mask attacks on 3DMask (see Fig. 6(a) in purple) with separate training spread out on both sides while...
Fig. 6. Feature distribution visualization. Best view in color.

Fig. 7. Average results on the joint face spoofing and forgery detection benchmark with different (a) compression bitrates; and (b) face resolutions.

the feature distributions clusters compactly with joint training (see Fig. 6(c) in purple). Similarly, it is clear to see from Figs. 6(b) and (d) that the feature distributions of different bonafides and attacks are likely to stand alone thus easier to find clear separation hyperplanes for distinguishing bonafide/attacks (e.g., DFDC_Bonafide in blue and CelebDFv2 in the purple of Fig. 6(d)). It demonstrates that joint training across two tasks enhances the models’ discrimination and generalization capacities for face attack detection.

In real-world scenarios, high video compression rate and low face resolution usually introduce serious compression artifacts and information losses. All these corruptions and quality degradations make the face spoofing and forgery detection challenging. Here we evaluate the performance under two challenging scenarios. First, we evaluate the appearance-based, rPPG-based, and appearance+rPPG based ViT(shared 8) on face spoofing and forgery detection under cross-testing scenarios with high compression rates (250/500/1000 kb/s) using x264 codec. The corresponding results are illustrated in Fig. 7(a). Compared with the appearance-based and appearance+rPPG based models, the performance of rPPG-based models degrade sharply when bitrates equal to 500 and 250 kb/s. This might be because the subtle rPPG clues are easily concealed by strong compression artifacts. Second, we evaluate the models with different low-resolution settings to mimic the long-distance video surveillance and low-resolution media scenarios. Specifically, bilinear interpolation is used to downsample the face frames to the sizes 16x16/32x32/64x64 first, and then upsample them back to the original face size. It is interesting to find that the corresponding results are illustrated in Fig. 7(b). Despite performance drops with lower face resolution for all models, the rPPG-based method is less sensitive to the resolution degradation compared with appearance-based method. The reason might be that the appearance-based method relies on the detailed texture info and contextual global semantics, which are heavily influenced by the resolution degradation. Overall, compared with appearance-based (easily influenced by resolution) and rPPG-based (easily influenced by video compression) methods, appearance+rPPG based method can explicitly mines complementary spoof/forgery clues, and achieves more robust performance when videos with different compression bitrates and resolutions.

VI. CONCLUSION

In this paper, we establish the first joint face spoofing and forgery detection benchmark using both visual appearance and physiological rPPG cues. We design a two-branch physiological network for discriminative rPPG cues mining, and the weighted normalization fusion strategy for efficient appearance+rPPG multi-modal models. We also investigate prevalent deep models, fusion strategies and joint training configurations on the proposed benchmark.

Findings: Extensive experiments were conducted on most of the proposed components. Here we summarize some interesting findings and issues learned from the extensive experiments: 1) the generalization capacities of both unimodal (appearance or rPPG) and multi-modal (appearance+rPPG) models can be obviously improved via joint training on both tasks. Specifically, two-branch physiological network is able to improve the performance of rPPG-based models while the weighted normalization fusion enhances the generalization and robustness (against compression and low-resolution distortions) of appearance+rPPG models. As for the settings of joint learning, training with 1 head 2 classes (bonafide/attack) and 2 heads 2 classes settings usually perform well while sharing the first 8 transformer blocks for joint training benefits generalization capacity. Finally, utilizing the simplest sampling strategy, i.e., randomly sampling minibatch data from two tasks is surprisingly effective; and 2) Despite improvement for generic visual models (e.g., ViT) in most datasets, joint training is not always effective for different models and datasets. For example, in Table II, in the case of unseen datasets Face Spoofing detection, the joint training is only 100% beneficial on the 3DMask dataset for all methods, whereas on the ROSE dataset for Xception, there is a drop of more than 10%in AUC. Similarly, for Mesonet on CelebDF. The reasons behind this might be i) the task-specific architectures cannot generalized well across task. For example, Xception usually performs well for deepfake task instead of anti-spoofing task. Thus, multitask
learning with spoofing data cannot guarantee the performance improvement; and ii) the gradient contradiction between these two tasks during the multitask learning. In other words, there will be some biases when training with data from multiple source domains for both two tasks. Thus, the trained models might perform well on specific datasets and tasks but poorly on some others.

**Future Works:** We note that the study of joint face spoofing and forgery detection is still at an early stage. Future directions include: 1) exploring more advanced multi-task learning and multi-modal fusion strategies; 2) besides rPPG-based color changes, exploiting semantic facial motion cues and contextual dynamic artifacts for joint face spoofing and forgery detection.

**REFERENCES**

[1] Y. Li, Y. Li, K. Xu, Q. Yan, and R. H. Deng, “Empirical study of face authentication systems under OSNFD attacks,” IEEE Trans. Dependable Secure Comput., vol. 15, no. 2, pp. 231–245, Mar./Apr. 2018.

[2] Z. Yu, Y. Qin, X. Li, C. Zhao, Z. Lei, and G. Zhao, “Deep learning for face anti-spoofing: A survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 45, no. 6, pp. 5609–5631, May 2023.

[3] J. Galbally and S. Marcel, “Face anti-spoofing based on general image quality assessment,” in Proc. 22nd Int. Conf. Pattern Recognit., 2014, pp. 1173–1178.

[4] X. Li, J. Komulainen, G. Zhao, P.-C. Yuen, and M. Pietikäinen, “Generalized face anti-spoofing by detecting pulse from face videos,” in Proc. 22nd Int. Conf. Pattern Recognit., 2016, pp. 4244–4249.

[5] Z. Yu, X. Li, and G. Zhao, “Facial-video-based physiological signal measurement: Recent advances and affective applications,” IEEE Signal Process. Mag., vol. 38, no. 6, pp. 50–58, Nov. 2021.

[6] J. Komulainen, A. Hadid, and M. Pietikäinen, “Context based face anti-spoofing,” in Proc. IEEE 6th Int. Conf. Biometrics: Theory Appl. Syst., 2013, pp. 1–8.

[7] K. Patel, H. Han, and A. K. Jain, “Secure face unlock: Spoof detection on smartphones,” IEEE Trans. Inf. Forensics Secur., vol. 11, no. 10, pp. 2268–2283, Oct. 2016.

[8] B. Lin, X. Li, Z. Yu, and G. Zhao, “Face liveness detection by rPPG features and contextual patch-based CNN,” in Proc. 3rd Int. Conf. Biometric Eng., 2019, pp. 61–68.

[9] S. Liu, P. C. Yuen, S. Zhang, and G. Zhao, “3D mask face anti-spoofing with remote photoplethysmography,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 85–100.

[10] S.-Q. Liu, X. Lan, and P. C. Yuen, “Multi-channel remote photoplethysmography correspondence feature for 3D mask face presentation attack detection,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 2683–2696, 2021.

[11] Y. Zhang et al., “Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 70–85.

[12] Y. Liu, J. Stehouwer, A. Jourabloo, and X. Liu, “Deep tree learning for zero-shot face anti-spoofing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2019, pp. 4680–4689.

[13] A. George, Z. Mostaani, D. Geissenbuhler, O. Nikisins, A. Anjos, and S. Marcel, “Biometric face presentation attack detection with multi-channel convolutional neural network,” IEEE Trans. Inf. Forensics Secur., vol. 15, pp. 42–55, 2019.

[14] A. Liu et al., “Contrastive context-aware learning for 3D high-fidelity mask face presentation attack detection,” IEEE Inf. Forensics Secur., vol. 17, pp. 2497–2507, 2022.

[15] Z. Yu et al., “Searching central difference convolutional networks for face anti-spoofing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 5295–5305.

[16] Z. Yu, X. Li, X. Niu, J. Shi, and G. Zhao, “Face anti-spoofing with human material perception,” in Proc. 16th Eur. Conf. Comput. Vis., Glasgow, U.K., Aug. 23–28, 2020, pp. 557–575.

[17] Y. Atoum, Y. Liu, A. Jourabloo, and X. Liu, “Face anti-spoofing using patch and depth-based CNNs,” in Proc. IEEE Int. Joint Conf. Biometrics, 2017, pp. 319–328.

[18] R. Cai, H. Li, S. Wang, C. Chen, and A. C. Kot, “DRL-FAS: A novel framework based on deep reinforcement learning for face anti-spoofing,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 937–951, 2020.
[44] Y. Liu, J. Stehouwer, and X. Liu, “On disentangling spoof trace for generic face anti-spoofing,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 406–422.

[45] R. Shao, X. Lan, J. Li, and P. C. Yuen, “Multi-adversarial discriminative deep domain generalization for face presentation attack detection,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 10023–10031.

[46] R. Shao, X. Lan, and P. C. Yuen, “Regularized fine-grained meta face anti-spoofing,” Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, pp. 11974–11981, 2020.

[47] R. Cai, Z. Li, R. Wan, H. Li, Y. Hu, and A. C. Kot, “Learning meta pattern for face anti-spoofing,” IEEE Trans. Inf. Forensics Secur., vol. 17, no. 12, pp. 2301–2312, 2022.

[48] Z. Li, H. Li, K.-Y. Lum, and A. C. Kot, “Unseen face presentation attack detection with hypersphere loss,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2020, pp. 2852–2856.

[49] O. Nikisins, A. Mohammadi, A. Anjos, and S. Marcel, “On effectiveness of anomaly detection approaches against unseen presentation attacks in face anti-spoofing,” in Proc. Int. Conf. Biometrics, 2018, pp. 75–81.

[50] Y. Qin, Z. Yu, L. Yan, Z. Wang, C. Zhao, and Z. Lei, “Meta-teacher for face anti-spoofing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 10, pp. 6311–6326, Oct. 2022.

[51] M. Rostami, L. Spinoulas, M. Hussein, J. Mathai, and W. Abd-Allameed, “Detection and continual learning of novel face presentation attacks,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 14851–14860.

[52] Z. Yu, Y. Shen, Z. Shi, H. Zhao, P. H. Torr, and G. Zhao, “Physformer: Facial video-based physiological measurement with temporal difference transformer,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 4186–4196.

[53] Z. Yu, W. Peng, X. Li, X. Hong, and G. Zhao, “Remote heart rate measurement from highly compressed facial videos: An end-to-end deep learning solution with frequency enhancement,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 151–160.

[54] E. M. Nowara, A. Sabharwal, and A. Veeraraghavan, “PPGSecure: Biometric presentation attack detection using photoplethysmograms,” in Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit., 2017, pp. 56–62.

[55] L. Birla, P. Gupta, and S. Kumar, “SUNRISE: Improving 3D mask face anti-spoofing for short videos using pre-emptive split and merge,” IEEE Trans. Dependable Secure Comput., vol. 20, no. 3, pp. 1927–1940, May/Jun. 2023.

[56] J. Yang, A. Li, S. Xiao, W. Lu, and X. Gao, “MTD-Net: Learning to detect deepfakes images by multi-scale texture difference,” IEEE Trans. Inf. Forensics Secur., vol. 16, pp. 4234–4245, 2021.

[57] C. Liu, H. Chen, T. Zhu, J. Zhang, and W. Zhou, “Making deepfakes more spurious: Evading deep face forgery detection via trace removal attack,” IEEE Trans. Dependable Secure Comput., vol. 20, no. 6, pp. 5182–5196, Nov./Dec. 2023.

[58] H. H. Nguyen, J. Yamagishi, and I. Echizen, “Capsule-forensics: Using capsule networks to detect forged images and videos,” in Proc. IEEE Int. Conf. Acoust. Speech Signal Process., 2019, pp. 2307–2311.

[59] A. Rossler, D. Cuzzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2018, pp. 7132–7141.

[60] S. Zhang et al., “A dataset and benchmark for large-scale multi-modal face anti-spoofing,” in Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit., 2019, pp. 919–928.

[61] Z. Yu, C. Zhao, K. H. Cheng, X. Cheng, and G. Zhao, “Flexible-modal face anti-spoofing: A benchmark,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2023, pp. 6345–6350.

[62] N. Erdogmus and S. Marcel, “Spoofing face recognition with 3D masks,” IEEE Trans. Inf. Forensics Secur., vol. 9, no. 7, pp. 1084–1097, Jul. 2014.

[63] D. Wen, H. Han, and A. K. Jain, “Face spoof detection with image distortion analysis,” IEEE Trans. Inf. Forensics Secur., vol. 10, no. 4, pp. 746–761, Apr. 2015.

[64] Z. Yu, J. Wan, Y. Qin, X. Li, Z. Li, and G. Zhao, “NAS-FAS: Static-dynamic difference network search for face anti-spoofing,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 1, pp. 3005–3023, Sep. 2021.

[65] H. Li, W. Li, H. Cao, S. Wang, F. Huang, and A. C. Kot, “Unsupervised domain adaptation for face anti-spoofing,” IEEE Trans. Inf. Forensics Secur., vol. 13, no. 7, pp. 1794–1809, Jul. 2018.

[66] B. Dohalnsky, R. Howes, B. Pfau, N. Baram, and C. C. Ferrer, “The deepfake detection challenge (DFDC) preview dataset,” 2019, arXiv:1910.08854.

[67] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celebr-DP: A large-scale challenging dataset for deepfake forensics,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2020, pp. 3207–3216.

[68] International organization for standardization, “ISO/IEC JTC1/SC37 biometrics: Information technology biometric presentation attack detection part 1: Framework,” 2016. [Online]. Available: https://www.iso.org/obp/ui/iso

[69] Z. Wang et al., “Domain generalization via shuffled style assembly for face anti-spoofing,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 4113–4123.

[70] S. Li and W. Deng, “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition,” IEEE Trans. Image Process., vol. 28, no. 1, pp. 356–370, Jan. 2019.

[71] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Process. Lett., vol. 23, no. 10, pp. 1409–1503, Oct. 2016.

[72] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770–778.
A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” in Proc. Int. Conf. Learn. Representations, 2021. [Online]. Available: https://openreview.net/forum?id=YicbFdNTTy

D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen, “MesoNet: A compact facial video forgery detection network,” in Proc. IEEE Int. Workshop Inf. Forensics Secur., 2018, pp. 1–7.

A. George and S. Marcel, “On the effectiveness of vision transformers for zero-shot face anti-spoofing,” in Proc. IEEE Int. Joint Conf. Biometrics (IJCB), 2021, pp. 1–8.

Y.-J. Heo, Y.-J. Choi, Y.-W. Lee, and B.-G. Kim, “Deepfake detection scheme based on vision transformer and distillation,” 2021, arXiv:2104.01353.

L. V. D. Maaten and G. Hinton, “Visualizing data using t-SNE,” J. Mach. Learn. Res., vol. 9, pp. 2579–2605, 2008.

Zitong Yu (Senior Member, IEEE) received the PhD degree from CMVS, University of Oulu, in 2022. He is an assistant professor with Great Bay University, China. He was a postdoctoral researcher with ROSE Lab, Nanyang Technological University. He was a visiting scholar with TVG, University of Oxford, from July to November 2021. His research interests include computer vision and biometric security. He was a recipient of IAPR Best Student Paper Award of CCBR 2023, IEEE Finland Section Best Student Conference Paper Award 2020, second prize of the IEEE Finland Jt. Chapter SP/CAS Best Paper Award (2022), and World’s Top 2% Scientists 2023 by Stanford.

Rizhao Cai received the BEng degree in electronic information engineering from Shenzhen University, China, in 2018. He is currently working toward the PhD degree with Nanyang Technological University, Singapore. He has been working as a project officer with the Rapid-Rich Object Search (ROSE) Laboratory and the NTU-PKU Joint Research Institute, leading projects that are collaborated with industrial partners. His research interests include computer vision and biometric security.

Zhi Li received the BEng degree in technology of detection and control from Shandong University, in 2017, the MSc degree in computer control and automation from Nanyang Technological University (NTU), Singapore, in 2018, and the PhD degree from the School of Computer Science and Engineering, in 2022. His research interests include machine learning, anomaly detection, and biometric security.

Wenhan Yang (Member, IEEE) received the BS and PhD degrees (Hons.) in computer science from Peking University, Beijing, China, in 2012 and 2018. He is currently an associate researcher with Peng Cheng Laboratory, Shenzhen, Guangdong, China. His current research interests include image/video processing/restoration, bad weather restoration, human-machine collaborative coding. He received the 2023 IEEE Multimedia Rising Star Runner-Up Award, the IEEE ICME-2020 Best Paper Award, the IJTC 2017 Best Paper Award, the IEEE CVPR-2018 UG2 Challenge First Runner-up Award, and the MSA-TC Best Paper Award of ISCAS 2022. He was the Candidate of CSIG Best Doctoral Dissertation Award in 2019. He served as the Area Chair of IEEE ICME-2021 and 2022, the Session Chair of IEEE ICME-2021, and the Organizer of IEEE CVPR-2019/2020/2021 UG2+ Challenge and Workshop.

Jingang Shi received the BS and PhD degrees from the Department of Electronics and Information Engineering, Xi’an Jiaotong University, China. From 2017 to 2020, he was a postdoctoral researcher with the Center for Machine Vision and Signal Analysis, University of Oulu, Finland. Since 2020, he has been an associate professor with the School of Software, Xi’an Jiaotong University. His current research interests mainly include image restoration, face analysis, and biomedical signal processing.

Alex C. Kot (Life Fellow, IEEE) was the head of the Division of Information Engineering and the vice dean research with the School of Electrical and Electronic Engineering. Subsequently, he was the associate dean with the College of Engineering for eight years. He has been with Nanyang Technological University, Singapore, since 1991. He is currently a professor and the director with the Rapid-Rich Object Search (ROSE) Laboratory and the NTU-PKU Joint Research Institute. He has published extensively in the areas of signal processing, biometrics, image forensics and security, computer vision, and machine learning. He is a fellow of the Academy of Engineering, Singapore. He received the Best Teacher of the Year Award and a coauthor of several best paper awards, including ICPR, IEEE WIFS, IEEE IWDW, CVPR Precognition Workshop, and VCIP. He has served as an associate editor for more than ten journals, mostly for IEEE Transactions. He served the IEEE SP Society in various capacities, such as the General Co-Chair for the 2004 IEEE International Conference on Image Processing and the vice-president for the IEEE Signal Processing Society. He was elected as the IEEE Distinguished Lecturer for the Signal Processing Society and the Circuits and Systems Society.