# 2D3MF - DEEPFAKE DETECTION USING MULTI-MODAL MIDDLE FUSION

*Adrian S. Roman, Aiden Chang, Hyunkeun Park, Kevin Hopkins, Shrutika Shrutika, Tom Yang*

Viterbi School of Engineering, University of Southern California, California, USA

## ABSTRACT

Deepfake detection is the task of detecting videos that have been generated or manipulated using deep learning. Detecting deepfakes is crucial to prevent the spread of misinformation in audio-visual media. Recent advancements in the field include joint learning of audio and visual information, by training independent modules and making a decision between learned embeddings from both modalities. While previous methods are robust when mainly the video content has been manipulated, they often face challenges when only the audio is manipulated. Our model, dubbed 2D3MF, proposes a novel method that exploits the relationship between emotions conveyed in audio and video for multi-modal deepfake detection. We benchmark our approach using RAVDESS, a popular emotion detection dataset, as well as three of the most commonly used deepfake datasets: DFDC, FaceForensics++, and FakeAVCeleb. Experiments demonstrate that the proposed framework using emotion embeddings followed by middle fusion, yields state-of-the-art performance and robust cross-dataset generalization on commonly used emotion and deepfake datasets.

Our code and models can be found at: `https://github.com/aiden200/2D3MF` and our project website can be found at `http://2d3mf.ahmen.io:20241/`.

## 1. INTRODUCTION

Recent progress in generative AI is steadily increasing the versatility and realism of deepfake video technology. In the audio domain, speech can be synthesized using text-to-speech (TTS) [1] or manipulated using voice conversion (VC) methods [2]. The video modality has also undergone advancements via generative algorithms such as Generative Adversarial Networks (GAN) [3], Difusion models [4], and Variational Autoencoders (VAE) [5]. With such deeplearning techniques, videos of real subjects, can be easily manipulated. This raises some ethical and legal concerns as the proliferation of modified media may be misused used to spread misinformation on the internet.

Various features have been proposed to perform deepfake detection [6, 7, 8]. Before the advent of advanced TTS or VC methods, most deepfake detection techniques were unimodal, focusing on identifying visual artifacts or traces from deepfake generation frameworks [9, 10]. However, these methods face challenges addressing modern manipulations that entail multiple modalities. Most recently, multimodal deepfake detection has particularly focused on detecting deepfakes in audio-visual data. Some methods target synchronization anomalies between audio and video [11, 12], flagging deepfakes when the two modalities are not time-aligned. Other methods extract embeddings from audio and video pre-trained models and make decisions by learning the relationship between the embeddings from either modality [7, 13]. Nevertheless, both approaches present shortcomings when only one modality is modified with high quality or when both modalities are corrupted.

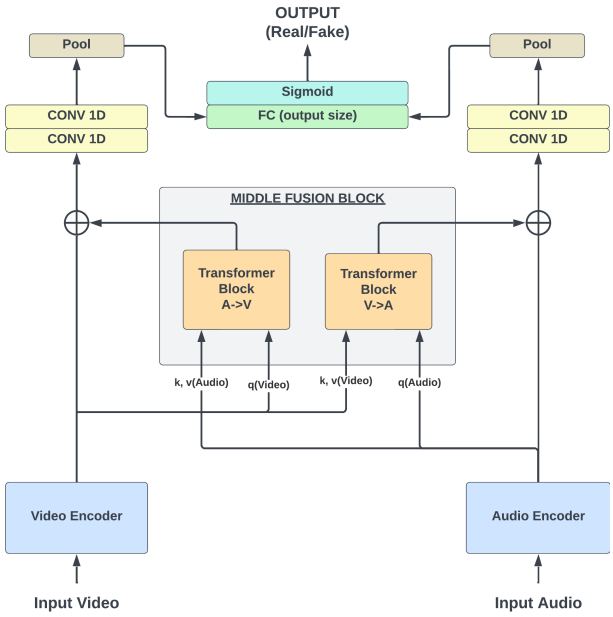


**Fig. 1**. Architecture of the 2D3MF model.

We hypothesize that by capturing emotional characteristics, the embeddings will contain rich, high-level and cross-modality features, that are well-suited for deepfake detection. We observe that the lack of naturality in deepfakes often stems from flaws in emotional expression, incongruous voice tones, and unnatural facial movements. To address these issues, we have developed various Deepfake Video Detection (DVD) architectures that integrate encoding models originally designed for tasks such as facial expression analysis, emotive analysis,

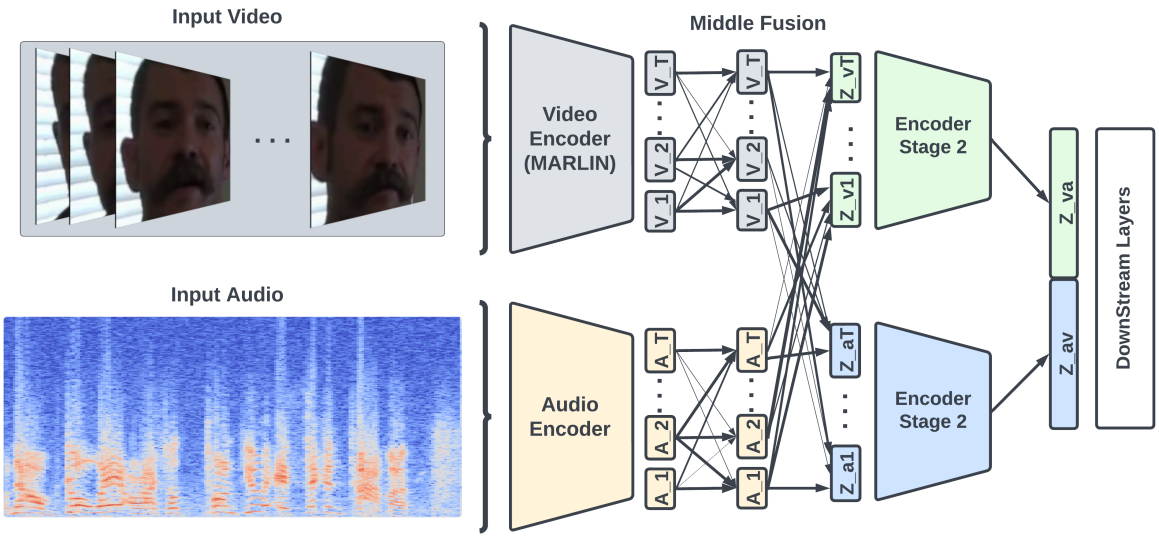


**Fig. 2**. High-level overview of our 2D3MF with audio and video inputs fused using Self and Cross-Attention middle fusion via transformer attention.

and speaker identification. These networks extract useful facial and speech representations, proving essential for the effectiveness of DVD tasks.

In this work, we study the utility of audio-visual emotion speaker embeddings, representations extracted from pre-trained audio and video networks, as robust features for the DVD task. To the best of our knowledge, we are the first to leverage abstract representations of emotions in the audio-visual domain to highlight and detect inconsistencies in fake videos. We propose 2D3MF (Deepfake Detection with Multi Modal Middle Fusion), which is a novel middle fusion strategy where audio and visual data are synergistically analyzed to capture discrepancies in emotional expressions, and vocal tones. These features reveal the subtle yet critical flaws inherent in deepfake videos, enabling our model to discern authentic content from manipulated media with precision.

In the video modality, we explore networks originally used as high level facial feature extractors. We use EfficientFace [14], which has been shown to work well on multi-modal emotion detection tasks [15]. We also integrate MARLIN [16], a video masked auto-encoder (MAE) that is capable of learning broad facial representations from videos. In the audio modality, we explore the use of features such as mel-frequency cepstral coefficients (MFCC), xvectors [17], pre-trained ResNet18 embeddings, emotion2vec embeddings [18], and audio masked autoencoder embeddings using EAT [19]. In summary, our contributions are:

1. We demonstrate that our framework, 2D3MF, produces embeddings that effectively capture emotional characteristics with minimal additional training, achieving state-of-the-art performance in classifying 8 different emotions.
2. Using this framework, we present a complete set of experiments that demonstrate the feasibility of audio and video emotion embeddings used for deepfake detection, achieving state-of-the-art performance on 3 different major deepfake datasets.
3. Finally we release a public GitHub repository where all the experiments presented can be easily reproduced.

## 2. RELATED WORK

### 2.1. Video Representation Learning

Most existing literature on deepfake detection is conducted in a supervised manner[16]. It has been shown that much of the state-of-the-art success in deepfake detection stems from accurately identifying high-level features[16]. However, achieving optimal results requires high-quality annotations and a large quantity of them. To address this issue, many recent models have adopted the approach of training using self-supervision, followed by fine-tuning in a supervised manner for domain specific downstream tasks. Among these self-supervised methods, Masked Auto Encoders (MAE) have been a popular example. In this approach, models are trained to reconstruct masked regions of a video using Vision Transformers [20]. Many of these works focus primarily on videos that feature human faces, learning to represent high-level facial features effectively.

### 2.2. Audio Representation Learning

In the audio domain, various methods of audio representation learning have been proposed. Two branches of represen-

tation learning can be identified: semi-supervised and self-supervised.

In the semi-supervised realm, previous research has shown that deep networks trained on large datasets learn to extract robust features that are invariant to resolution, domain shifts, or media artifacts. These features, obtained from intermediate outputs of deep networks are referred to in the literature as 'embeddings', and they are often used in downstream tasks. In this study we leverage the work of Aldeneh et al. [21] that uses xvector embeddings extracted from a speaker identification network. In their research, xvectors were found to be robust at abstracting emotions features, which we then show to be robust for the downstream task of deepfake detection. Following the utility of emotion embeddings for deepfake detection tasks, we pre-train a ResNet18 network to perform 8 emotion classification using the RAVDESS dataset. We extract embeddings from this pre-trained network.

In the self-supervised realm, recent work has introduced models such as emotion2vec [18], which uniquely performs universal speech emotion learning. Similar to the video domain, audio Masked Auto Encoders (MAEs) have emerged as powerful architectures. By learning to reconstruct masked spectrograms, they capture both local and global information about the audio data. This enables audio MAEs, such as audioMAE [22] and EAT [19], to extract relevant high-level speech information. We initially employ this in emotion detection, as discussed in section 4.1, and subsequently apply it to deepfake detection, as detailed in section 4.2).

### 2.3. Multi-modal Representation Learning

Multi-modal approaches can be classified into three categories. The first is early fusion, a popular technique where inputs from multiple modalities are concatenated before being fed as an input to the model. This approach may not be suitable for modalities with different characteristics, as features from one modality could dominate the others. Methods such as [11, 12] have applied early fusion to analyze features such as synchronization. The second category, late fusion, involves learning embeddings for each modality separately and concatenating them at the end for downstream tasks. Prior works like [12] use late fusion, although these methods struggle to effectively learn embeddings from both modalities. Our architecture employs the third category, middle fusion. Middle fusion applies cross-attention to the learned embeddings before a second-stage encoder processes them, effectively learning from both modalities. IAT [23] demonstrated the effectiveness of middle fusion in emotion detection.

## 3. METHODOLOGY

### 3.1. Datasets

We use a total of 5 audio-visual datasets in this study: DFDC [24], Faceforensics++ [25], FakeAVCeleb [**?**], DeepfakeTIMIT [**?**] and RAVDESS [26]. These datasets feature clips where the speaker is facing the camera.

The DeepFake Detection Challenge Dataset (DFDC) is the largest video deepfake dataset containing more than 100,000 video clips. We filtered the dataset to keep only videos with a single subject in front of the camera. Since this study makes use of other datasets, to maintain data balance with other datasets used in this study, we retained only 5,200 video clips. Faceforensics++ comprises 5000 video clips collected from YouTube and manipulated using methods such as: Deepfakes [27], NeuralTextures [28], Face2Face [29], and FaceSwap [30]. We filtered the Faceforensics++ dataset to contain clips with only one active subject. After filtering, and due to YouTube data availability, our final dataset contains 1,400 video clips. FakeAVCeleb consists of 500 real videos and more than 20,000 fake videos from five ethnic groups, each with 100 real videos from 100 subjects. We consider four categories within this dataset: fakeAudio+fakeVideo (FAFV), fakeAudio+realVideo (FARV), realAudio+fakeVideo (RAFV), and realAudio+realVideo (RARV). DeepfakeTIMIT consists of 320 fake videos from 32 subjects manipulated using the FaceSwap technique. We use the higher-quality data in this study. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) contains over 1,000 utterances and high-resolution videos for 24 actors showcasing different emotions.

We follow the train/eval/test splits specified by each dataset. For training our model to identify deepfakes, we use DeepfakeTIMIT, which contains only fake data, and the RAVDESS dataset, which contains only real data, exclusively for training to ensure a balanced representation of fake and real content. Our training, validation, and test sets all contain both 'real' and 'fake' video samples, labeled as 0 or 1, respectively. For emotion detection, we benchmark our model using the RAVDESS dataset, classifying eight different emotions. We follow the train/test/val splits provided in the RAVDESS repo.

### 3.2. Metrics

To assess the performance of our model and baselines, we use classification accuracy (ACC) and the Area Under the Curve (AUC), as these metrics commonly benchmark state-of-the-art models. We focus on ACC when benchmarking emotion detection due to the presence of multiple classes, while we emphasize AUC for deepfake detection, given the highly unbalanced nature of the dataset.

### 3.3. Video Encoder

Our literature review indicates that models trained on low-level video features tend to underperform when analyzing low-quality videos, including those that have been modified through processes like compression. In contrast, models like LipForensics [31] prioritize learning lexical content by focusing on features in the mouth region. Building on this approach, we plan to extend the scope of learning to include not only the high-level features around the mouth area but also the overall facial structure.

To facilitate broader learning, we propose employing random masking on portions of the image (effectively setting their values to zero). This method will encourage the model to recognize and learn a diverse range of features, thereby enhancing its adaptability and accuracy. With this in mind, we leverage MARLIN [16] as the feature extractor in the video modality. MARLIN excels in its understanding of facial features—ranging from the nose to the eyes—through an advanced encoder-decoder architecture. Furthermore, MARLIN exhibits impressive performances across a variety of downstream tasks. Its applications extend to emotion recognition and the identification of deepfake content, showcasing its embedding's adaptability. We have modified the MARLIN model to enable it to capture more compact information across extended timesteps, thereby enhancing the temporal dimension of each video.

### 3.4. Audio Encoder

The characteristics of the audio encoder differ significantly from those of the video modality. Our goal is to study audio representations, specifically embeddings from networks, that abstract relevant information about speakers with little variability. As a result, we explore a variety of audio backbones. We leverage five distinct audio encoders:

- **MFCC:** we extract 10 Melfrequency cepstral coefficients (MFCC) per second. These are simple features that we adopt as our baseline as previous literature shows their robustness for speech related tasks.
- **Xvectors:** embeddings extracted from a deep neural network. Aladeh et al., [21] demonstrated the relevance of xvectors embeddings for emotion recognition tasks.
- **ResNet18:** following a similar apprach as with xvectors, we experimented with a Resnet18 architecture pre-trained to classify 8 emotions from the RAVDESS dataset. We use the intermediate outputs from the second to last layer as out embeddings.
- **emotion2vec:** with the same relationship between the robustness of emotion embeddings for deepfake detection, we adopt the use of the emotion2vec [18] network, a self-supervised network that generates emotion embeddings from raw audio waveforms.
- **EAT [19] (Efficient Audio Transformer):** an Audio Masked Auto Encoder (MAE), excels at capturing both local and global audio characteristics. It generates a high amount of temporal dimensions, achieving state-of-the-art (SOTA) results in tasks such as environmental sound classification, speaker identification and speech command recognition. In subsequent sections, we demonstrate how effectively this capability translates to emotion classification and, subsequently, to deepfake detection.

### 3.5. 2D3MF

We construct the 2D3MF model to capture high-level emotional characteristics, such as robust facial feature detection which we anticipate will translate effectively to deepfake detection. Hence, our proposal model follows a similar architecture to a study that performs multimodal emotion recognition architecture, initially designed for emotion detection, to effectively analyze both video and audio cues.

Our video and audio embeddings are extracted per second using our respective encoders, as detailed in sections 3.3 and 3.4. These embeddings are subsequently fused through cross-attention transformer blocks, creating a multimodal representation. This representation leverages characteristics of visual and auditory information to capture temporal dependencies and implement attention mechanisms. The fused data is then processed by a second-stage encoder that learns from both modalities. The architecture of our transformer block is illustrated in Figure 3. During our ablation studies, we modify the transformer block to either refuse or more effectively capture characteristics within a single modality.

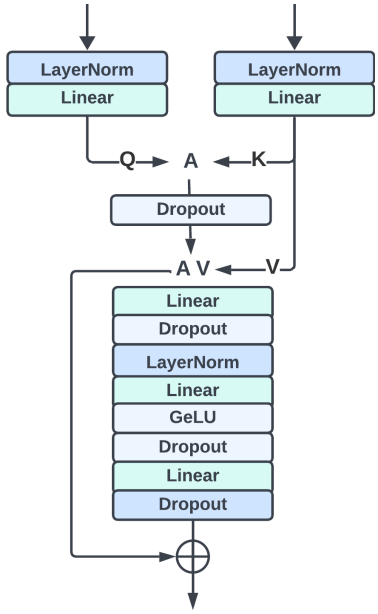


**Fig. 3**. Architecture of the transformer block

Let $z_a$ and $z_v$ represent the embeddings generated by the first stage encoders for the audio and visual modalities, respectively. In each modality's processing branch, a transformer block is incorporated. For this example, within the audio branch, the transformer block uses the visual features, $z_v$, to generate keys and values, while queries are derived from the audio features, $z_a$. Attention, or $A$ in our transformer block diagram, is calculated as the following:

$$A = \text{softmax}\left(\frac{z_a W_q (W_k z_v)^T}{\sqrt{d_v}}\right) z_v W_v,$$

where $W_q$, $W_k$, and $W_v$ are the weight matrices for queries, keys, and values, respectively, and $d_v$ is the dimensionality of the visual features. We have outlined the diagram of our model 2D3MF in figure 1.

## 4. RESULTS

| Model | ACC ↑ |
|---|---|
| Marlin Small | 0.46 |
| Marlin + xvectors | 0.558 |
| Marlin + emotion2vec | 0.571 |
| EAT | 0.639 |
| IAT [23] | 0.815 |
| Marlin + EAT | **0.836** |

**Table 1**. Performances comparison on test set from RAVDESS. Classifying 8 different emotions.

| Model | Modality | ACC ↑ | AUC ↑ |
|---|---|---|---|
| Marlin Base (baseline) | VO | .77 | .836 |
| Marlin + MFCC | AV | 0.856 | 0.851 |
| Marlin + xvectors | AV | 0.9516 | 0.9823 |
| Marlin + ResNet18 | AV | 0.9449 | 0.9825 |
| Marlin + emotion2vec | AV | 0.9081 | 0.962 |
| Marlin + EAT | AV | **0.9533** | **0.9897** |
| RealForensics [32] | AV | N/A | 0.956 |

**Table 2**. Performances comparison on test set from FaceForensics++. The SOTA method is shown at the bottom.

### 4.1. Emotion recognition performance

We first benchmark our model on the RAVDESS dataset to assess how effectively our model's embeddings can generalize to emotion recognition. We tested the model on eight distinct emotions without training on any additional data. Our results were compared to the Intermediate Attention Fusion (IAT) model [23], which is recognized as the current state-of-the-art on RAVDESS without additional data training. Notably, IAT is ranked second overall; however, we excluded the top-ranked model from our comparisons due to its training on a larger dataset, for fair comparison. The results are detailed in Table 1, where our model demonstrated state-of-the-art performance in emotion classification achieving a 83.6% ACC. Particularly, the combination of EAT and Marlin performed the best. This performance is attributed to EAT's capability to effectively capture both local and global audio information and its ability to present a high amount of temporal dimensions. We would like to mention that while IAT trimmed its video inputs to 3.6 seconds to capture essential video information efficiently, our approach standardized video inputs to an average duration of 10 seconds, which is the average time across all the datasets we used in our experiments (including the deepfake datasets). This standardization resulted in almost all videos being padded, introducing non-essential audio data into the analysis.

| Model | Modality | ACC ↑ | AUC ↑ |
|---|---|---|---|
| Marlin Base (baseline) | VO | .79 | .842 |
| Marlin + MFCC | AV | 0.8144 | 0.8924 |
| Marlin + xvectors | AV | 0.9323 | 0.9677 |
| Marlin + ResNet18 | AV | 0.9372 | 0.975 |
| Marlin + emotion2vec | AV | 0.9246 | 0.971 |
| Marlin + EAT | AV | **0.9497** | **0.9885** |
| Face X-ray [33] | AV | N/A | 0.9347 |

**Table 3**. Deepfake performances comparison on test set from DFDC. The SOTA method is shown at the bottom.

| Model | Modality | ACC ↑ | AUC ↑ |
|---|---|---|---|
| Marlin + MFCC | AV | 0.6317 | 0.8546 |
| Marlin + xvectors | AV | 0.9032 | 0.9585 |
| Marlin + ResNet18 | AV | 0.9429 | 0.9671 |
| Marlin + emotion2vec | AV | **0.9642** | 0.9181 |
| Marlin + EAT | AV | 0.9442 | **0.9868** |
| AVFakeNet [34] | AV | 0.9259 | 0.92 |

**Table 4**. Performances comparison on test set from FakeAVCeleb. The SOTA method is shown at the bottom.

### 4.2. Deepfake detection performance

Table 2 shows the performance of our proposed models against one of the state-of-the-art (SOTA) models on FaceForensics++. We present the baseline performance as the unimodal, video-only (VO) MARLIN Base model, which achieves 0.836 AUC and 0.77 ACC. On the other hand, using MFCCs as audio features, which makes the architecture audio-visual (AV), allows for performance improvements. This showcases the potential of using an audio-visual archi-

tecture for deepfake detection. Further performance improvements are achieved with more sophisticated audio feature extractors. On this test dataset, xvector and ResNet18 seem to achieve similar performance when included in the audio-visual task. On the other hand, the benefits of emotion2vec compared to the baseline are clear; however, other audio backbones appear to be more robust in terms of their AUC and ACC scores. The best performing model, Marlin + EAT, achieves the highest AUC (0.9897) and ACC (0.9533), outperforming the SOTA model, Realforensics [32], by 0.05% in AUC.

When it comes to the DFDC dataset, we observe that all of our proposed architectures outperform the baseline model (see Table 3). Interestingly, for this benchmark, the emotion2vec audio backbone contributes to better deepfake detection performance compared to xvectors or ResNet18. We attribute this trend to the higher quality of video forgeries in DFDC, where emotion2vec more effectively detects relationships with the audio domain, thereby improving classifications. Overall, DFDC is a larger test set that contains a greater variety in deepfake forgeries. This variety also contributes to the slight performance degradation observed in Marlin + EAT. Nevertheless, most of our proposed architectures outperform the state-of-the-art model (Face X-ray [33]).

We also performed benchmarking on the FakeAVCeleb dataset (see Table 4). This dataset is particularly relevant because it contains a higher proportion of audio-based forgeries than other datasets. As a result, our benchmarking of different audio backbones shows more variability. This time, emotion2vec and MFCCs produced suboptimal results, indicating that these methods may not be effective and cannot generalize well to scenarios where the audio is predominantly modified. However, other methods such as xvectors and ResNet18 prove resilient under such conditions. Overall, the best-performing method we achieved is Marlin+EAT, which also outperforms the state-of-the-art model (AVFakeNet [34]) by 0.06% in AUC and 0.02% in ACC.

Note: all experiments presented in this section used the train splits from 5 datasets from section 3.1, evaluated on the validation split and tested on the test split of the dataset of interest.

### 4.3. Deepfake cross-dataset generalization

When it comes to deepfake datasets, it is common for the train, evaluation, and test sets to all contain similar forgery techniques. This can be a source of concern when the intention is to deploy a deepfake detection model for real-world applications. For this reason, we test the generalization capabilities of our best model, Marlin + EAT, using unseen datasets. Table 4.3 showcases our model's generalization capabilities, having been trained on various datasets but tested on different, unseen ones. For these experiments, we use all five datasets presented in section 3.1, training on four and leaving one for testing. Thus, we demonstrate our model's generalization across five different conditions, each time selecting a different dataset's test split for benchmarking. Overall, we observe that our model's classification capabilities regress by only 0.02% on average compared to the original benchmarks presented in Tables 2, 3, and 4. It is also worth noting that Marlin + EAT performs quite well in classifying real-only videos (using RAVDESS as test set) and fake-only videos (using DeepfakeTIMIT as test set).

### 4.4. Ablation Studies

For our ablation studies, three primary approaches were employed. The first was dropping out a modality during training. Previous research [23] has indicated that introducing random noise or dropping a modality can prevent the model from overly depending on a single modality, thus enhancing performance in scenarios where one modality may be compromised or absent. However, our models exhibited decreased performance under these conditions, showing approximately 0.1 lower scores in the AUC metric for deepfake detection. We believe this is due to the nature of our datasets, where all but one contain only one modality that is falsified. As such, depending on the dropped modality, our model's ability to differentiate between fake and real media will be significantly impeded.

The second ablation study focused on different types of fusion techniques. As our pretrained first-stage encoders were not originally designed to process multiple modalities simultaneously, we evaluated our model by comparing middle fusion with late fusion, omitting the second-stage encoders for late fusion. During late fusion, we observed an average decrease in performance of approximately 10% in AUC, affecting both deepfake detection and emotion recognition tasks.

The third ablation study was conducted using different middle fusion blocks, where we experimented with combining audio and video modalities in six different ways. The default fusion mechanism is detailed in Section 3.5. In this context, $z_a$ and $z_v$ denote the embeddings generated by the first-stage encoders for the audio and visual modalities, respectively. The transformer block $T_{1,2}$, illustrated in Figure 3, employs modalities 1 and 2 as the key/value and query modalities, respectively. The symbol $h$ represents the output of the transformer block $T$. Each middle fusion approach manipulates these embeddings differently:

- **Default Middle Fusion:**

$$h_{av} = T_{av}(z_v, z_a), \quad h_{va} = T_{va}(z_a, z_v)$$

- **Audio Reuse Fusion:**

$$h_{va} = T_{va}(z_a, z_v), \quad h_{av} = T_{av}(z_v + h_{va}, z_a)$$

- **Video Reuse Fusion:**

$$h_{av} = T_{av}(z_v, z_a), \quad h_{va} = T_{va}(z_a + h_{av}, z_v)$$

| Train datasets | Test datasets | ACC ↑ | AUC ↑ |
|---|---|---|---|
| All but DFDC | DFDC | 0.935 | 0.978 |
| All but FF++ | FF++ | 0.932 | 0.961 |
| All but FakeAVCeleb | FakeAVCeleb | 0.907 | 0.979 |
| All but RAVDESS | RAVDESS | 0.9917 | 0.961 |
| All but DeepFakeTIMIT | DeepFakeTIMIT | 0.896 | 0.934 |

**Table 5**. Cross-dataset generalization on our best performing architecture: MARLIN+EAT

- **Self-Attention Fusion:**

$$h_{va} = T_{vv}(z_v, z_v), \quad h_{av} = T_{aa}(z_a, z_a)$$

- **Multi-Attention Fusion:**

$$\begin{aligned} h_{av1} = T_{av1}(z_v, z_a), \quad h_{va1} = T_{va1}(z_a, z_v), \\ z_a = z_a + h_{av1}, \quad z_v = z_v + h_{va1}, \\ h_{av2} = T_{av2}(z_v, z_a), \quad h_{va2} = T_{va2}(z_a, z_v), \\ h_{av} = h_{av2}, \quad h_{va} = h_{va2} \end{aligned}$$

- **Self and Cross-Attention Fusion:**

$$\begin{aligned} h_{vv} = T_{vv}(z_v, z_v), \quad h_{aa} = T_{aa}(z_a, z_a), \\ z_v = z_v + h_{vv}, \quad z_a = z_a + h_{aa}, \\ h_{av} = T_{av}(z_v, z_a), \quad h_{va} = T_{va}(z_a, z_v) \end{aligned}$$

We evaluated our different middle fusion techniques on both RAVDESS for an emotion benchmark and DFDC for a deepfake benchmark. We use Marlin + EAT, as the combination of these two models outperformed all other models. We show our results on table 6.

For emotion recognition performance, we see that our default middle fusion block (one layer of cross attention transformer) performed the best. We had hypothesised that due to the high performance in audio only over video, the audio modality would play a more crucial factor into our transformer block, but did not see much improvement on audio refuse. Since our first stage encoders have self-attention built in them, we observed that our default fusion, where we performed cross attention once, performed the best. We think this is because that our first stage encoders have strong representations of each modality individually, and due to the small dataset of RAVDESS, it was most effective when only applying transformer block once.

For deepfake detection performance, we observed a marginal change of less than 1% in performance metrics. Such a minor variation is not statistically significant, given the fluctuations in AUC performance across different runs. We attribute this stability to the amount of data utilized for training, which likely provided a robust foundation for the model, diminishing the impact of small modifications in the processing pipeline.

| Model | Task | ACC ↑ | AUC |
|---|---|---|---|
| Video Refuse | Emotion | 0.596 | 0.936 |
| Audio Reuse | Emotion | 0.679 | 0.942 |
| Self & Cross-Attention | Emotion | 0.712 | 0.948 |
| Multi-Attention | Emotion | 0.741 | **0.962** |
| Self-Attention | Emotion | 0.779 | 0.956 |
| Default | Emotion | **0.836** | 0.954 |

**Table 6**. Ablation studies on different middle fusion techniques.

## 5. CONCLUSION

We have successfully demonstrated that our 2D3MF architecture excels in emotion recognition, translating this proficiency into deepfake detection. Throughout our experiments, our framework has proven its ability to leverage emotional characteristics extracted from pre-trained audio and video networks to accurately identify manipulated content. By integrating facial feature extractors such as EfficientFace and video masked auto-encoder (MAE) technologies like MARLIN, coupled with different audio feature extractors, our approach significantly enhances deepfake detection. The use of multimodal data not only overcomes the limitations of previous detection methods but also significantly improves detection accuracy, even in scenarios where one or both modalities have been falsified.

### 5.1. Future Directions

In the next phase of our project, we aim to enhance the usability of our software by packaging and distributing our codebase via PyPI. This will simplify the installation process for users, hopefully increasing the feedback from the research community.

Additionally, we plan to conduct extensive experiments focusing on hyperparameter optimization to refine the models' performance further. We are also considering the integration of additional emotional datasets to strengthen the applicability and robustness of our framework.

Our ultimate goal is to submit our work to the International Conference on Acoustics, Speech, and Signal Process-

ing (ICASSP) by August 2024. This will allow us to showcase our findings to a global audience and contribute meaningful advancements to the field of speech and signal processing. We plan to compress this paper to four pages; therefore, we have written the minimum amount of pages (8 pages) for this current report to prepare for theses requirements.

### 5.2. Future Work

We encourage further exploration of our work into different modalities, integrating embeddings within those modalities that capture emotional characteristics. Our code is publicly available, designed for easy integration with additional audio-visual backbones if desired.

## 6. ACKNOWLEDGEMENTS

The authors thank Prof. Yue Zhao for his support and guidance throughout the semester. We would also like to thank TA Pengda Xiang for the enriching midterm project conversations and feeback. Finally we would like to thank CARC at USC for providing HPC resources for this study.

## 7. REFERENCES

[1] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu, "A survey on neural speech synthesis," *arXiv preprint arXiv:2106.15561*, 2021.

[2] Yishuang Ning, Sheng He, Zhiyong Wu, Chunxiao Xing, and Liang-Jie Zhang, "A review of deep learning based speech synthesis," *Applied Sciences*, vol. 9, no. 19, 2019.

[3] 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*, vol. 63, no. 11, pp. 139–144, 2020.

[4] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang, "Diffusion models: A comprehensive survey of methods and applications," *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.

[5] Jaehyeon Kim, Jungil Kong, and Juhee Son, "Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5530–5540.

[6] Gereon Fox, Wentao Liu, Hyeongwoo Kim, Hans-Peter Seidel, Mohamed Elgharib, and Christian Theobalt, "Videoforensicshq: Detecting high-quality manipulated face videos," in *2021 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2021, pp. 1–6.

[7] Hasam Khalid and Simon S Woo, "Oc-fakedect: Classifying deepfakes using one-class variational autoencoder," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 656–657.

[8] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung, "Deepfake detection: A systematic literature review," *IEEE access*, vol. 10, pp. 25494–25513, 2022.

[9] Zhixi Cai, Shreya Ghosh, Abhinav Dhall, Tom Gedeon, Kalin Stefanov, and Munawar Hayat, "Glitch in the matrix: A large scale benchmark for content driven audio–visual forgery detection and localization," *Computer Vision and Image Understanding*, vol. 236, pp. 103818, 2023.

[10] Weifeng Liu, Tianyi She, Jiawei Liu, Run Wang, Dongyu Yao, and Ziyou Liang, "Lips are lying: Spotting the temporal inconsistency between audio and visual in lip-syncing deepfakes," *arXiv preprint arXiv:2401.15668*, 2024.

[11] Chao Feng, Ziyang Chen, and Andrew Owens, "Self-supervised video forensics by audio-visual anomaly detection," 2023.

[12] Sahibzada Adil Shahzad, Ammarah Hashmi, Yan-Tsung Peng, Yu Tsao, and Hsin-Min Wang, "Av-lip-sync+: Leveraging av-hubert to exploit multimodal inconsistency for video deepfake detection," *arXiv preprint arXiv:2311.02733*, 2023.

[13] Heqing Zou, Meng Shen, Yuchen Hu, Chen Chen, Eng Siong Chng, and Deepu Rajan, "Cross-modality and within-modality regularization for audio-visual deepfake detection," *arXiv preprint arXiv:2401.05746*, 2024.

[14] Zengqun Zhao, Qingshan Liu, and Feng Zhou, "Robust lightweight facial expression recognition network with label distribution training," in *Proceedings of the AAAI conference on artificial intelligence*, 2021, vol. 35, pp. 3510–3519.

[15] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj, "Self-attention fusion for audiovisual emotion recognition with incomplete data," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 2822–2828.

[16] Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and

Munawar Hayat, "Marlin: Masked autoencoder for facial video representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2023, pp. 1493–1504, IEEE.

[17] David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[18] Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen, "emotion2vec: Self-supervised pre-training for speech emotion representation," *arXiv preprint arXiv:2312.15185*, 2023.

[19] Wenxi Chen, Yuzhe Liang, Ziyang Ma, Zhisheng Zheng, and Xie Chen, "Eat: Self-supervised pre-training with efficient audio transformer," 2024.

[20] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas Zamir, Fahad Shahbaz Khan, and Mubarak Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[21] Zakaria Aldeneh and Emily Mower Provost, "You're not you when you're angry: Robust emotion features emerge by recognizing speakers," *IEEE Transactions on Affective Computing*, vol. 14, no. 2, pp. 1351–1362, 2021.

[22] Po-Yao Huang, Hu Xu, Juncheng Li, Alexei Baevski, Michael Auli, Wojciech Galuba, Florian Metze, and Christoph Feichtenhofer, "Masked autoencoders that listen," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28708–28720, 2022.

[23] Kateryna Chumachenko, Alexandros Iosifidis, and Moncef Gabbouj, "Self-attention fusion for audiovisual emotion recognition with incomplete data," *CoRR*, vol. abs/2201.11095, 2022.

[24] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer, "The deepfake detection challenge (dfdc) dataset," *arXiv preprint arXiv:2006.07397*, 2020.

[25] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner, "Faceforensics++: Learning to detect manipulated facial images," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.

[26] Steven R Livingstone and Frank A Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, pp. e0196391, 2018.

[27] TT Nguyen, CM Nguyen, DT Nguyen, DT Nguyen, and S Nahavandi, "Deep learning for deepfakes creation and detection. arxiv 2019," *arXiv preprint arXiv:1909.11573*, 2019.

[28] Justus Thies, Michael Zollhöfer, and Matthias Nießner, "Deferred neural rendering: Image synthesis using neural textures," *Acm Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.

[29] Hyeongwoo Kim, Mohamed Elgharib, Michael Zollhöfer, Hans-Peter Seidel, Thabo Beeler, Christian Richardt, and Christian Theobalt, "Neural style-preserving visual dubbing," *ACM Transactions on Graphics (TOG)*, vol. 38, no. 6, pp. 1–13, 2019.

[30] Michał Zendran and Andrzej Rusiecki, "Swapping face images with generative neural networks for deepfake technology–experimental study," *Procedia computer science*, vol. 192, pp. 834–843, 2021.

[31] Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic, "Lips don't lie: A generalisable and robust approach to face forgery detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.

[32] Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic, "Leveraging real talking faces via self-supervision for robust forgery detection," 2022.

[33] 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*, 2020, pp. 5001–5010.

[34] Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik, "Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection," *Applied Soft Computing*, vol. 136, pp. 110124, 2023.