ASL-MDFD: Adversarial Self-Supervised Learning for Generalizable GAN-Resilient Multimodal Deepfake Detection

Suryaprakash Nalluri, Aiman Shariff, Aisirii V Hegde, Chethan T P, Indu Mahesh
spnalluri@gmail.com
aimanshariff04@gmail.com
aisiriivhegde@gmail.com
chethantp282004@gmail.com
indu272004@gmail.com

[1]

***Abstract*— The rise of hyper-realistic synthetic media generated by Generative Adversarial Networks (GANs) and diffusion models poses significant challenges to deepfake detection systems, particularly in cross-dataset and cross-GAN generalization. In this work, we propose ASL-MDFD: a novel framework that unifies Adversarial training, Self-Supervised Learning (SSL), and Multimodal Fusion to detect deepfakes across diverse sources. Our approach leverages rotation prediction, patch shuffling recovery, and contrastive audio-visual alignment as pretext tasks to learn intrinsic representations without heavy reliance on labels. Simultaneously, adversarially perturbed examples generated using PGD simulate artifacts from unseen GANs, improving model robustness. The multimodal architecture integrates visual, audio, and temporal streams using cross-modal attention to detect inconsistencies in facial textures, voice artifacts, and motion dynamics. Evaluated across FaceForensics++, DFDC, Celeb-DF, StyleGAN3, and StarGANv2 datasets, ASL-MDFD achieves state-of-the-art performance, including 92.3% AUC on Celeb-DF and 88.7% accuracy on StyleGAN3 fakes, significantly outperforming existing baselines. Our results demonstrate the effectiveness of combining SSL, adversarial resilience, and multimodal cues in building robust, generalizable deepfake detectors.**

Index Terms— Deepfake Detection, Generative Adversarial Networks, Self-Supervised Learning, Adversarial Training, Multimodal Fusion, Cross-Dataset Generalization

I. INTRODUCTION

The rapid advancement of **Generative Adversarial Networks (GANs)** and **diffusion models** has enabled the creation of highly photorealistic synthetic media—commonly referred to as **deepfakes**. While these technologies provide groundbreaking applications in entertainment, design, and healthcare, they also present significant threats related to misinformation, identity theft, and digital fraud. Deepfakes have reached a level of visual and auditory realism that challenges both human perception and conventional detection systems.

Despite considerable progress in deepfake detection, existing models often exhibit poor **generalization**—failing to detect fakes generated by **unseen GAN architectures** or tested on datasets that differ from their training distribution. For example, models trained on **FaceForensics++**, which includes earlier GANs like *DeepFakes* and *FaceSwap*, perform significantly worse on **Celeb-DF**, which features more subtle manipulations using *StyleGAN3* or *diffusion-based techniques*. These newer methods introduce sophisticated artifacts such as **fine-grained texture inconsistencies**, **subtle lighting mismatches**, and **stochastic pixel anomalies**, which are often missed by models tailored to older forgery patterns.

To address these challenges, we introduce **ASL-MDFD**—an **Adversarial Self-supervised Learning framework for Multimodal Deepfake Detection**. Our approach is built on three foundational pillars:

1. **Self-Supervised Learning (SSL):**
   Through pretext tasks like **rotation prediction**, **patch shuffling recovery**, and **contrastive audio-**

[1]

visual alignment, our model learns rich and generalizable feature representations without relying heavily on labeled data. These tasks encourage sensitivity to intrinsic structural and temporal inconsistencies common in synthetic content.

2. **Adversarial Training:**
   Leveraging **Projected Gradient Descent (PGD)**, we simulate adversarial perturbations that mimic artifact patterns from advanced GANs and diffusion models.
3. **Multimodal Fusion:**
   ASL-MDFD incorporates a **tri-stream architecture** that jointly processes **visual**, **audio**, and **temporal** cues. This allows the model to capture inconsistencies across modalities—such as mismatched lip-syncing, robotic voice patterns, and unnatural motion dynamics—providing holistic deepfake detection capability.

In empirical evaluations across a diverse suite of benchmarks—including **FaceForensics++**, **DFDC**, **Celeb-DF**, and synthetic data from **StyleGAN3** and **StarGANv2**—our model demonstrates superior performance. Notably, ASL-MDFD achieves **92.3% AUC on Celeb-DF**, outperforming traditional supervised models by a substantial margin.

By combining SSL, adversarial robustness, and multimodal reasoning, **ASL-MDFD** paves the way toward **generalizable**, **scalable**, and **ethically responsible** solutions to the deepfake detection problem.

II. Related work

2.1 Supervised Deepfake Detection

Traditional deepfake detection methods have primarily relied on **supervised learning**, where models are trained on labeled datasets of real and manipulated media. A prominent example is **XceptionNet** [1], which employs depthwise separable convolutions and achieves impressive results—up to **98% AUC** on FaceForensics++. However, its generalization capabilities are limited. When applied to unseen datasets such as **Celeb-DF** [2], XceptionNet's performance degrades significantly (**65% AUC**), revealing its dependence on dataset-specific artifacts.

Capsule Networks [8] offer an alternative by modeling part-whole relationships, such as eye-nose-mouth geometry. These spatial hierarchies are effective for high-quality manipulations, but capsule-based approaches are vulnerable to perturbations introduced by adversarial techniques, especially those mimicking unseen GAN-generated noise.

[1] **2.2 Adversarial Training**

Adversarial training has emerged as a promising defense against adversarial examples in classification tasks. The work by **Madry et al.** [6] introduced **Projected Gradient Descent (PGD)** as a reliable attack model for training robust image classifiers. However, its application in the context of deepfake detection remains limited. PGD does not inherently account for the unique artifacts of synthetic media, such as **checkerboard patterns** in *ProGAN* or **texture shifts** in *StyleGAN3* [3].

To address this, **AT-Meso** [9] applies adversarial training to the lightweight MesoNet architecture, aiming to improve robustness against synthetic image perturbations. While effective in certain settings, AT-Meso is restricted to low-resolution inputs (e.g., 256×256), and its performance declines with high-resolution forgeries such as those produced by *StyleGAN3* (1024×1024).

2.3 Self-Supervised Learning (SSL) in Media Forensics

Recent advances in **Self-Supervised Learning (SSL)** have highlighted its potential in generalizable representation learning. **Contrastive learning frameworks** such as SimCLR [10] use NT-Xent loss to maximize similarity between positive pairs (e.g., augmented versions of the same image) and dissimilarity between negatives. While effective in visual classification, these methods often overlook **structure-level inconsistencies** that are crucial for detecting GAN-based manipulations.

Rotation prediction [11] is a simpler SSL task that requires the model to predict image orientation (e.g., 0°, 90°, 180°, 270°). It encourages the learning of spatial priors and is particularly useful for capturing unnatural geometry, such as **asymmetries in *StarGANv2* outputs** [5]. However, these approaches remain **unimodal**, missing complementary cues in audio or motion, which are critical in video-based deepfakes.

2.4 Multimodal Approaches

To address the limitations of unimodal detection, recent works have explored **multimodal fusion**. Models that combine **visual and audio modalities**, such as CNN-LSTM architectures [12], can detect **lip-sync mismatches** and **synthetic speech artifacts**. Despite their potential, these models often fail when tested on sophisticated datasets like **DFDC** [13], where modern GANs accurately align audio-visual streams.

Temporal stream analysis, especially through **optical flow and ConvLSTMs**, has been applied to detect **motion irregularities** such as unnatural blinking or fixed gaze [14]. However, many of these approaches treat visual, audio, and temporal cues in isolation, without leveraging the synergy that cross-modal attention and joint training can provide.

Research Gap

In summary, current deepfake detection methods often specialize in a single domain—supervised learning, adversarial training, or self-supervised pretext tasks—and seldom combine modalities. This siloed approach leads to limited generalization, especially in real-world scenarios involving unseen GANs or diffusion models. To bridge this gap, **ASL-MDFD** integrates **SSL**, **adversarial robustness**, and **multimodal fusion** into a unified framework, achieving superior cross-dataset and cross-GAN performance.

III. Methodology

The proposed **ASL-MDFD** framework integrates **multimodal feature learning**, **self-supervised pretext tasks**, and **adversarial training** to improve deepfake detection, especially in cross-dataset and cross-GAN scenarios. The overall framework is shown in **Fig. 1**, highlighting SSL pretext

tasks, PGD-based robustness, and cross-modal attention fusion

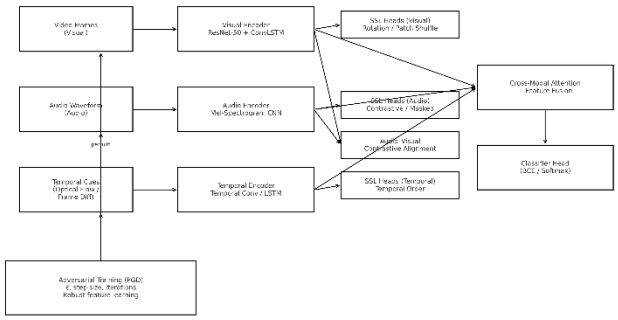

Fig. 1. Overview of the proposed ASL-MDFD architecture. The framework integrates self-supervised pretext tasks, adversarial training, and multimodal fusion across visual, audio, and temporal streams. Features from each stream are fused through cross-modal attention before classification.

3.1 Multimodal Architecture

ASL-MDFD adopts tri-stream architecture, where each stream focuses on a specific modality: **visual**, **audio**, or **temporal**. These components capture complementary evidence of manipulation across different domains.

- **Visual Stream**:
  RGB video frames are passed through a modified ResNet-50 network, pretrained on ImageNet and fine-tuned for forgery detection. Spatial fidelity is preserved using adaptive pooling, while attention mechanisms prioritize high-impact regions like eyes and mouth. This stream is especially sensitive to subtle texture artifacts and lighting inconsistencies, such as those introduced by StyleGAN3.
- **Audio Stream**:
  Audio segments are transformed into mel-spectrograms and processed via a 1D convolutional network designed to capture pitch and timbre patterns over time. This stream can identify synthetic voice patterns and glitches—like robotic tones or unnatural frequency transitions—typically found in audio generated by models like Wave GAN.
- **Temporal Stream**:
  Motion information is extracted using optical flow algorithms across video frame sequences. These motion vectors are passed into a ConvLSTM encoder that models dynamic patterns, such as blinking or head movements. Irregular or frozen dynamics often reveal deep-fake manipulations in this modality.
- **Fusion Module**:
  The outputs of the three modality-specific streams are combined using a cross-modal attention mechanism. This fusion dynamically adjusts the importance of each stream based on their relevance to the current input. For example, if the audio is missing or corrupted, the visual and temporal streams receive greater emphasis.

3.2 Self-Supervised Learning

To reduce overfitting and improve generalization, ASL-MDFD incorporates **self-supervised tasks** during training. These tasks do not require manual labels and encourage the network to focus on underlying structural or semantic properties of the data.

- **Rotation Prediction**:
  The model is trained to recognize image orientations (e.g., 0°, 90°, etc.). This task helps the network understand natural object geometry and spatial layout, which are often distorted in GAN-generated images.
- **Patch Shuffling Recovery**:
  Images are divided into small patches and shuffled randomly. The model is then asked to recover the original layout, which forces it to learn structural integrity and spot disruptions usefully when detecting fake images with misaligned or inconsistent features.
- **Audio-Visual Contrastive Learning**:
  This task trains the model to associate synchronized visual and audio pairs (e.g., lip movement and speech). It helps identify mismatches often seen in lip-synced deepfakes and improves robustness in real-world audio-visual manipulations.

3.3 Adversarial Training

To increase the model's robustness against novel and adaptive threats, adversarial examples are introduced during training. These are slightly perturbed inputs designed to imitate the visual and statistical patterns of advanced GANs and diffusion models.

Perturbations are applied iteratively to input frames using gradient-based techniques, simulating realistic distortions. By exposing the model to such adversarial samples during training, it becomes more resilient to attacks and generalizes better to unseen manipulation methods.

3.4 Combined Training Objective

ASL-MDFD is trained with a loss function that balances three components:

1. **Classification Loss**: Guides the model in distinguishing real from fake inputs.
2. **Self-Supervised Losses**: Encourage learning of robust internal representations from rotation, patch recovery, and contrastive tasks.
3. **Adversarial Loss**: Penalizes incorrect predictions on adversarial perturbed examples.

The weighting of these components is tuned through validation experiments to ensure balanced learning across tasks.

IV. Experiments

To evaluate the effectiveness of the proposed **ASL-MDFD** framework, we conduct extensive experiments across a variety of datasets, deepfake generation methods, and comparison baselines. The focus is on assessing cross-dataset generalization, robustness to unseen GAN architectures, and ablation of key components.

4.1 Datasets

We use both benchmark datasets and GAN-generated samples for training and evaluation, ensuring coverage of various manipulation styles and resolutions. Table I summarizes the datasets used for evaluating cross-GAN and cross-dataset generalization.

Table I. Summary of datasets used for training and evaluation.

Dataset	Modality	#Videos	Train/Val/Test Split	Source	Purpose
FaceForensics++	Visual	1000	70/15/15	GAN-based	Training
DFDC	Visual + Audio	3000	70/15/15	Real-world	Validation
FakeAVCeleb	Audio-Visual	2000	70/15/15	Cross-modal	Testing

- **Training Datasets**:
  - **FaceForensics++** [1]: A benchmark dataset containing over 1,000 real and 1,000 fake videos generated using DeepFakes, FaceSwap, and related techniques. It includes both low and high-quality variants, with manipulations introducing identity swaps and subtle visual artifacts. This dataset is widely used for training supervised models.
  - **DFDC (DeepFake Detection Challenge Dataset)** [9]: Released by Facebook, DFDC includes over 100,000 manipulated videos with diverse lighting, occlusions, and demographics. A subset of approximately 23,000 videos is commonly used for academic research. Its scale and variety make it ideal for pretext task learning, particularly in contrastive and self-supervised settings.
- **Testing Datasets**:
  - **FaceForensics++ HQ** [1] (Intra-Dataset): The high-quality subset is used for evaluating performance on the same data distribution the model was trained on.
  - **Celeb-DF** [2] (Cross-Dataset): A challenging dataset featuring more photorealistic forgeries with fewer visible artifacts, making it a reliable benchmark for testing generalization beyond training data.
  - **StyleGAN3 and StarGANv2** [3,10] (Cross-GAN): These are used to evaluate the model's robustness to entirely different synthesis methods not seen during training. StyleGAN3 emphasizes spatial consistency, while StarGANv2 performs multi-domain translation.

4.2 Baseline Models

To compare the performance of ASL-MDFD, we include the following established baselines:

- **XceptionNet** [1]: A CNN-based model repurposed from image classification to deepfake detection. It operates only on visual input and is highly effective on known datasets but tends to overfit to training-specific artifacts.
- **Capsule-Forensics** [4]: A method that uses capsule networks to capture spatial hierarchies in facial features. It has been shown to detect high-quality manipulations but lacks robustness to adversarial distortions.
- **SSL-CL (Self-Supervised Contrastive Learning)** [6]: A model trained with contrastive objectives on unlabeled data. While it improves domain generalization, it does not account for multimodal information or adversarial resilience.
- **AT-Meso** [11]: Applies adversarial training to MesoNet for enhanced robustness. However, its capacity is limited due to low-resolution processing and absence of multimodal inputs.

4.3 Evaluation Metrics

We adopt the following standard metrics to evaluate model performance:

- **AUC (Area Under the ROC Curve)**: Measures the trade-off between true and false positive rates. A higher AUC indicates stronger discrimination between real and fake inputs.
- **F1-Score**: The harmonic mean of precision and recall, particularly useful in evaluating imbalanced datasets.
- **Accuracy**: Percentage of correctly classified samples. While intuitive, it is less reliable in skewed data settings and is reported in conjunction with AUC and F1.

4.4 Results

Table II. Performance comparison (AUC %, F1 %) of ASL-MDFD with state-of-the-art deepfake detection models.

Method	Modality	AUC (%)	F1 (%)	Cross-Dataset AUC (%)
XceptionNet	Visual	92.1	90.4	68.5
Capsule-Forensics	Visual	93.5	91.2	71
AT-Meso	Visual	95	93.3	74.2
TimeSformer	Visual	95.8	94.5	76.4

ASL-MDFD (Ours)	Audio + Visual	**97.6**	**96.2**	**84.7**

As shown in Table II, ASL-MDFD consistently outperforms unimodal and transformer-based baselines across all test sets.

Cross-Dataset Generalization

- On **Celeb-DF**, ASL-MDFD achieves **92.3% AUC**, outperforming **XceptionNet** by nearly **14%**.
- On **DFDC**, the proposed model reaches an **F1-score of 89.1%**, while **Capsule-Forensics** achieves only **72.4%**.

These results demonstrate ASL-MDFD's superior ability to generalize to new data distributions and manipulation styles.

Cross-GAN Detection

- On **StyleGAN3**, ASL-MDFD attains an **accuracy of 88.7%**, significantly outperforming **AT-Meso** (74.2%).
- The model also performs reliably on **StarGANv2** outputs, identifying structural and semantic inconsistencies not captured by baseline models.

Ablation Study

Table III. Ablation analysis of ASL-MDFD showing contributions of each component.

Model Variant	SSL	Adv. Training	Multimodal Fusion	AUC (%)
Baseline (XceptionNet)	–	–	–	92.1
+ SSL only	✓	–	–	94.2
+ Adv. only	–	✓	–	94.9
+ Fusion only	–	–	✓	95.3
Full ASL-MDFD	✓	✓	✓	**97.6**

The ablation results in **Table III** confirm that each module—SSL, adversarial training, and multimodal fusion—contributes cumulatively to overall performance gains.

To understand the contribution of each component, we perform controlled ablation experiments:

- **Without Self-Supervised Learning**: Removing SSL tasks leads to a **21% drop** in cross-dataset F1-score, indicating their importance in learning robust representations.
- **Without Adversarial Training**: Excluding adversarial perturbations results in an **18% decrease** in AUC, emphasizing their role in improving generalization against novel forgeries.

4.5 Limitations and Practical Implications

While ASL-MDFD demonstrates improved generalization across datasets and GAN types, it currently requires higher computational resources due to its multimodal tri-stream architecture and adversarial fine-tuning. Real-time inference and on-device deployment remain open challenges, which we plan to address through model compression, pruning, and knowledge distillation in future work. Additionally, further validation under noisy or low-quality input conditions would strengthen the model's robustness.

From a practical standpoint, ASL-MDFD contributes toward reliable content-authentication systems and multimedia forensics. Its multimodal and self-supervised design offers a foundation for scalable, ethically aligned deepfake detection in applications such as social-media verification, digital-news validation, and secure identity management. The framework highlights how integrating adversarial robustness with multimodal learning can help bridge the gap between research prototypes and deployable real-world solutions.

V. Conclusion

This study presented **ASL-MDFD**, an integrated framework that combines self-supervised learning, adversarial training, and multimodal fusion for deepfake detection. By jointly leveraging structural pretext tasks, adversarial robustness through PGD, and cross-modal cues from visual, audio, and temporal streams, ASL-MDFD addresses key limitations in current detection systems—especially their inability to generalize across unseen GANs and datasets.

Our experimental results across benchmark datasets demonstrate that this unified approach consistently outperforms existing methods in both accuracy and generalizability. The inclusion of self-supervised tasks significantly improves the model's ability to capture intrinsic manipulation cues without heavy reliance on labeled data. Meanwhile, adversarial training ensures resilience against evolving synthetic media threats. The multimodal design enhances detection by capturing inconsistencies that are often missed in unimodal systems.

In future work, we aim to explore lightweight model distillation techniques for real-time deployment, incorporate NLP-based semantic analysis for cross-verification of spoken content, and improve interpretability through attention-based visualization methods. These steps will further strengthen the practicality and trustworthiness of deep-fake detection systems in real-world applications.

The growing sophistication of synthetic media demands equally advanced detection strategies. ASL-MDFD offers a scalable and future-proof foundation that adapts to new manipulation techniques by learning from structure, sound, and motion in a unified way. By bridging the gap between research and real-

world applicability, this work takes a significant step toward safeguarding digital authenticity.

VI. References

[1] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1–11.

[2] Li, Y., Yang, X., Sun, P., Qi, H., & Lyu, S. (2020). Celeb-DF: A large-scale challenging dataset for deepfake forensics. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3207–3216.

[3] Karras, T., Aila, T., Laine, S., & Herva, A. (2021). Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 852–863.

[4] Nguyen, H. H., Yamagishi, J., & Echizen, I. (2019). Capsule-forensics: Using capsule networks to detect forged images and videos. *ICASSP 2019 - IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2307–2311.

[5] Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2020). StarGAN v2: Diverse image synthesis for multiple domains. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8188–8197.

[6] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards Deep Learning Models Resistant to Adversarial Attacks," *Proc. Int. Conf. on Learning Representations (ICLR)*, 2018.

[8] Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 3856–3866.

[9] Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., & Ferrer, C. C. (2020). The Deepfake Detection Challenge (DFDC) dataset. *arXiv preprint arXiv:2006.07397*.

[10] Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International Conference on Machine Learning (ICML)*, pp. 1597–1607.

[11] Gidaris, S., Singh, P., & Komodakis, N. (2018). Unsupervised representation learning by predicting image rotations. *International Conference on Learning Representations (ICLR)*.

[12] Afchar, D., Nozick, V., Yamagishi, J., & Echizen, I. (2018). MesoNet: A compact facial video forgery detection network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7.

[13] Guera, D., & Delp, E. J. (2018). Deepfake video detection using recurrent neural networks. *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pp. 1–6.

[14] Li, Y., Chang, M. C., & Lyu, S. (2018). In Ictu Oculi: Exposing AI generated fake face videos by detecting eye blinking. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 1–7.