INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 7, July 2026 | ISSN: 2456-4184 | IJNRD.ORG

Next-Gen Deepfake Detection: Harnessing Graph Transformers

B. V. Wakode

Assistant Professor, Information Technology Department, GCOE Amravati, India

Abstract : The rapid advancement of deepfake technology has produced totally unreal fake media. In this paper a comprehensive study is presented on how to create a deepfake and how to detect. The study is presented with three aspects; how the deepfake is created, how invariant risk minimization applies to detecting a deepfake and how a self-supervised graph transformer can be used to detect deepfakes. The first part of the study was done by reviewing many of the past researched papers focused on how machine learning and deep learning are being used to create deepfakes. The second part of the study used Invariant Domain-oriented(ID3) Deepfake Detection Methodology, to generate a method for improving performances of deepfake detection across domains by using Domain Refinement, Data Augmentation and Domain Calibration. The final aspect included building a new framework for detecting deepfakes using a self-supervised pre-training model to combine vision Transformers and graph convolution networks with a Transformer discriminator for improved accuracy and explainability. A diverse range of experimental tests concluding the ID3 methodology and the new framework showed that both methods generalize well and are resilient against perturbations that occur during post-processing of deepfake video.

IndexTerms **- Deepfake, Machine Learning, Deep Learning**

I. INTRODUCTION

Rapid advances in artificial intelligence (AI) and deep learning have given rise to highly realistic synthetic media, often referred to as deepfakes. These AI-generated alterations to images and videos are capable of convincingly portraying people in situations or performing actions that never occurred. The emergence of such technologies raises serious concerns with respect to privacy, security and misinformation [1]. In addition as the technology mature it is becoming increasingly difficult to differentiate between real and altered content. For this reason, developing a reliable and efficient means of detecting deepfakes is imperative to preserve the integrity of digital media. One of the primary challenges currently facing deepfake detection methods is that they generally do not generalize between datasets or types of manipulations. Furthermore, many models use supervised learning where the model is trained on a large labeled dataset. As a result, these models will usually produce acceptable results when tested on data that is similar to their training data; however, they will struggle to produce acceptable results when tested on previously unseen or unknown data. As such, many of the methods currently available for detecting deep fakes are limited due to their ability to detect deep-fake media, which may have been produced using varied methods of production and have undergone post-processing modifications. The above limitations clearly demonstrate the need for approaches for detecting deepfakes that are more robust and capable of achieving higher detection rates across a broader range of scenarios. Self-supervised learning techniques are another way of improving the effectiveness of deep fake detection algorithms [2]. These algorithms learn from unlabelled data, allowing for more robust feature extraction. Using contrastive learning, a model can learn to better extract high-level visual features that are less sensitive to features such as compression or other types of post-processing. This reduces the chance that the high-level visual features extracted will be unique to one specific dataset so that the model will be able to generalise the results when exposed to a new type of deep fake. A graph convolutional network (GCN) can be integrated with a transformer-based discriminator to help the model better understand complex relationships between different parts of an image.

Another approach aimed at enhancing the generalizability of deepfake detection models employs IRM (invariant risk minimization). By seeking to maximize the amount of information that can be learned by any model in general rather than only learning from the trained set this approach allows for greater image representation and identity differentiation. Therefore, the IRM-assisted deepfake detection system will provide greater robustness towards instances of deepfakes that were constructed with entirely different techniques or datasets than used for the training sample.

Integrating self-supervised learning, graph-based models and invariant risk minimization can yield a comprehensive framework for detecting deepfakes [3]. This new approach has the potential to overcome several limitations associated with current systems, including overfitting and inability to recognize post-processing artifacts. It also provides a more effective and adaptable means of detecting deepfakes across a wider array of situations. As the threat posed by deepfakes increases it is increasingly important that detection methods continue to be refined and improved to maintain the integrity and accuracy of digital media.

II. LITERATURE SURVEY

As a result of the rapid growth of deepfake technology through the use of sophisticated artificial intelligence along with the use of deep learning algorithms in order to create artificially manipulated media. There has been a significant increase in the amount of study being conducted on how to accurately and effectively detect media that has been manipulated using deepfake technologies. Most traditional approaches for detecting deepfake media have largely focused on identifying low-level visual artifacts, such as those resulting from texture inconsistencies and blending anomalies through convolutional neural networks (CNNs) that detect common artifacts that are a result of the deepfake generation process. These traditional methods are very effective at detecting deepfake media under controlled conditions [4]. However, they are not always able to accurately or effectively handle media that has undergone various forms of post-processing. Advancements in the detection area are focusing on the development of better and more generalizable solution comparatives to identify items of interest. A promising method for accomplishing this is to utilize self-

supervised or self-taught learning techniques. This type of learning eliminates the need for a large amount of labelled data because it uses the input data itself as the source of the label. In accomplishing this high level visual features can be extracted that are relatively independent of a specific set of data as well as being more resilient to common forms of corruption. Vision transformers built using self-supervised or self-taught contrastive learning to pre-train their feature extraction capabilities have demonstrated considerable potential to identify relationships between these features in order to enable the detection of subtle manipulations that traditional detection techniques would likely miss.

A different innovative method consists of combining graph-based models with existing types of neural network architectures. Facial images can be accurately represented as graphs where the nodes correspond to the facial landmarks and edges correspond to the spatial relationships between these areas or regions of the face. As a result, this type of graph representation is capable of providing a more detailed analysis of areas where faces are manipulated and therefore make it much more effective when analyzing deepfakes. The use of graph convolutional networks (GCNs) along with transformers has been successful in representing both local and global dependencies within such graphs leading to improved detection abilities [5].

The use of Invariant Risk Minimization (IRM) is also increasingly being adopted for purposes of improving deepfake detection model generalization performance [6]. Conventional models frequently overfit training data and thus have reduced performance when evaluated on new, unseen data. IRM seeks to remedy this by training models to discover invariant features across multiple environments or datasets, which diminishes the dependence of the IRM-trained model on dataset characteristics [7]. This technique has successfully improved the adaptability of detection models so they can more readily detect deepfakes that have been generated using many different methods or from many separate datasets. Moreover, deepfake detection models are becoming more and more reliant on being able to explain their predictions [8]. The ability to explain why models arrive at certain decisions is vital for building trust and being transparent with users. Graph transformer-based models with relevance maps provide visualizations to help users see/reason about which parts of an image were most impactful for making a model's prediction. These types of models should promote transparency as well as be able to point out what features correspond to deepfake manipulation.

The area of deepfake detection is rapidly developing. Current practices are moving to more advanced techniques that provide better generalization, robustness and explainability. Examples would include The use of self-supervised learning, graph based models and invariant risk minimization; Signifies that there is improvement in developing systems for detecting deepfakes that can be utilised in the ever-evolving and complex world of deepfake technologies.

III. PROPOSED METHODOLOGY

The framework for detecting deepfakes provides advanced approaches to improve the generalization and robustness of the detection system such as self-supervised learning, graph based approaches and invariant risk minimization. The framework is built to overcome the limitations of classic detection algorithms which can often fail when confronted with unseen deepfakes or altered media due to post processing.

A) Self-Supervised Feature Extraction:

The large number of labelled training examples needed to train a deep learning algorithm is often a major obstacle in the creation of robust deepfake detectors. To overcome the challenge of not having many labelled examples available for developing a deepfake detection algorithm, the proposed self-supervised learning framework will utilize contrastive learning as its self-supervised learning technique. Contrastive learning provides a way for the proposed framework to learn representations of unlabelled examples, by contrasting two augmented views of the same unlabelled image with all other unlabelled examples in the data source.

The self-supervised learning method adopted by the proposed framework allows for maximum flexibility of the resultant deepfake detector because the features learned are high-level visual representations that do not depend upon the data sets used for training. As a result, the resultant deepfake detector will have a higher capacity for detecting deepfakes from multiple manipulation types, including those not seen during training. Further, because the resultant deepfake detector is trained on general features, it will have better robustness to post-processing artefacts (e.g. compression, noise) that are frequently used to hide deepfake impostures.

Contrastive learning has been used to make sure that the system can manage many different forms of deformation, which helps the detection system's overall generalisation ability. By training the model in a self-supervised fashion, there is less demand for large labelled datasets and the model can then be successfully applied in a much wider range of use cases.

B) Graph-Based Detection for Facial Manipulations

Manipulations of facial areas, like the eyes, mouth and nose, are frequently minor in nature when creating a deepfake. To capture those changes adequately, our proposed method uses a graph representation of the face, where key facial landmarks are represented by nodes and the relationship between those landmarks is represented by edges.

When using this graph representation of the face, we will have a more exact representation of the face which enables us to do more detailed analysis of the structure of the face. The Graph Convolutional Network (GCN), which will process the graph, will find local changes by examining the relationship between feature points on the face; this is especially important in finding deepfakes with only some portions of the face altered (e.g., partial manipulation, facial swap).

In addition, GCN will work in conjunction with a transformer classifier, which will enhance detection ability since it examines global relationships between the face's feature points. The GCN will identify changes at the local level (e.g., a minor change to the eyes), while the transformer will identify the overall patterns of the feature points on the face. This combined approach allows for the identification of localized and global changes at a better rate than if only one approach was used alone. Therefore, using this model, we will be able to detect deepfakes at all levels, including very fine (small) and very complex (ghost) deepfakes.

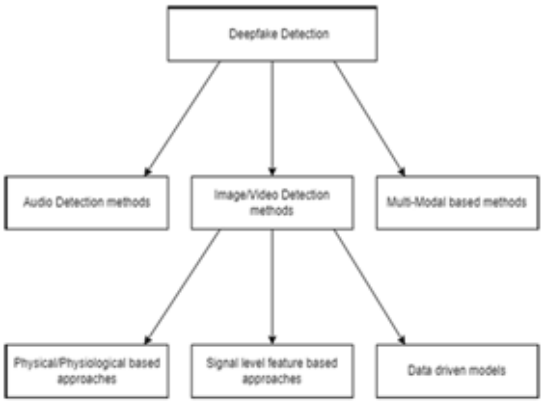
The framework increases the ability to detect localized alterations that may not be seen with conventional CNN-based methods, by utilizing both the GCN and transformers to describe the face as a graph. This allows for a more detailed representation of facial landmarks, which will result in greater robustness against detecting facial structure changes.

C) Invariant Risk Minimization (IRM) for Generalization

Many deepfake detector systems struggle to generalize well onto other datasets due to the variability of deepfake examples based upon unknown generations. Traditional model types will typically "overfit" and perform poorly on unseen (or more broadly defined) datasets. Consequently, Invariant Risk Minimization (IRM) was incorporated into the training regime to help address these challenges.Through techniques including Invariant Representation Matching (IRM), a model can be trained on characteristics that are steady across several sets or conditions. Rather than establishing characteristics solely based on the data within the training set, the model learns to understand all sets of the same invariant feature resulting in it being significantly more capable of generalizing across distinctive deepfake types and data, thus further positioning itself within an industry that is continuously changing due to advancements in deepfake technology.

The detection system can still achieve high performance when faced with previously unseen deepfake generation methods through the use of IRM. This ensures that the feature set learned from the data used to train the model will be capable of predicting true class membership in other contexts and will thus perform effectively when deployed in the real world, where deepfakes exist in differing degrees of quality, using different post-processing techniques and/or different creation methods.

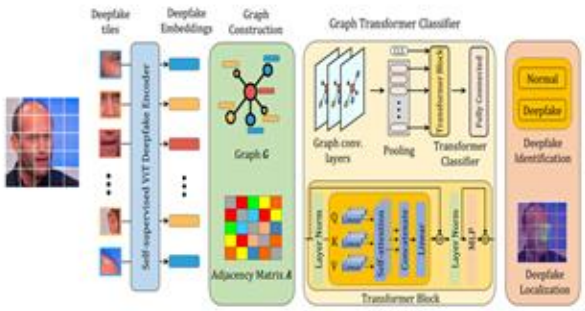
Additionally, this technique will prevent overfitting to the spurious relationships found within specific datasets and improve the generalization capability of the model, thus enabling it to perform effectively on a range of different data sources. Consequently, this will help to keep the detection system incomparably robust, despite the continuous evolution of the methods used to create deepfakes.



D) Explainability Through Relevance Maps

Although it is important to obtain high accuracy when detection occurs, it is equally critical that decisions made by the system will be transparent and interpretable. One way in which the interpretability of the system's decision is achieved by the use of the framework's relevance map created through an attention mechanism within the graph transformer model, which gives insight into how the system arrives at its decisions.The relevance maps identify areas of the face that influenced the decision the model made. It also visually indicates which areas of the face were coded as "important" in determining whether the picture being classified was real or fake. For instance, if a deepfake modified only the eye region of the face, the relevance map would identify these areas as the most important when determining the model's final prediction.

The explainability functionality of the system enhances user trust as it provides an understanding for the operators of the reasons behind certain decisions. Furthermore, it also assists with debugging and improving the model as a developer can use the relevance map to ascertain the areas of focus when the model may have been incorrect, therefore continuing to enhance its accuracy over time. Attention mechanisms integrated into a transformer provide a fully interpretable way for discerning how the model arrives at its decision, facilitating the process of validating its predictions. The resulting level of interpretability establishes trust in the system, allowing it to be employed in sensitive use cases in which it is important to know the reasoning behind a particular decision.



.

IV. CONCLUSION

The proposed framework for detecting fake videos (fake media) uses Self-Supervised Learning (SSL), Graph-Based Models and Invariant Risk Minimization (IRM) paired with relevance maps to provide an interpretable and reliable solution. Each component promotes the benefits of providing a solution to traditional methods that do not generally offer much in the way of robustness or interpretability. By employing SSL, we can build a model that does not require an extensive labelled training data set nor does it require robust operation under various distortion factors like compression. By mapping the input video data into a graph-based model, we can focus on specific, discrete manipulations of facial characteristics to create distinct representations for multiple types of fake media. In the final step of creating the model, the IRM component ensures that there is at least one representation that is focused on all potential variations of traditional fake media and any new types of fake media that may be developed in the future. The result of combining SSL, Graph-Based Models and IRM with relevance maps is a framework that is not only interpretable but also enhances the transparency and reliability of the framework.

REFERENCES

[1] R. Tolosana, M. Á. Aguera, A. Montalvo, and J. Fierrez, "DeepFakes: A New Threat to Face Recognition", IEEE Access, vol. 8, pp. 24061–24070, 2020, doi: 10.1109/ACCESS.2020.2972955.
[2] Z. Liu, Y. Yang, and Y. Xie, "DeepFake Detection Based on Deep Learning: A Survey", IEEE Transactions on Information Forensics and Security, vol. 15, pp. 3053–3068, 2020, doi: 10.1109/TIFS.2020.2982214.
[3] R. Ranjan and R. Chellappa, "Hands-on Guide to Face Manipulation: Data, API, and Challenges", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019.
[4] M. A. Abdelrahman and A. H. Akl, "Real-Time Deepfake Detection Using Transfer Learning," Journal of Real-Time Image Processing, 2022.
[5] Ming-Yu Liu, Xun Huang, Jiahui Yu, Ting-Chun Wang, and Arun Mallya, "Generative adversarial networks for image and video synthesis: Algorithms and applications", Proceedings of the IEEE, 109(5):839–862, 2021.
[6] C. Chan, S. Ginosar, T. Zhou, and A. A. Efros, "Everybody Dance Now", Proceedings of the IEEE International Conference on Computer Vision, 2019.
[7] K. M. Malik, A. Javed, H. Malik, and A. Irtaza, "A light-weight replay detection framework for voice controlled IOT devices," IEEE Journal of Selected Topics in Signal Processing, 2020.
[8] Rana MS, Nobi MN, Murali B, Sung AH, "Deepfake detection: A systematic literature review", IEEE access 2022.Feb 24;10:25494-513.
[9] Rafique R, Gantassi R, Amin R, Frnda J, Mustapha A, Alshehri AH. Deep fake detection and classification using error-level analysis and deep learning. Scientific reports. 2023.
[10] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan, "Recurrent convolutional strategies for face manipulation detection in videos", Interface 2019.