AI-Based Forensic Face Sketch Generation and Suspect Identification from Eyewitness Verbal Description

Manikanda Prabhu V[1], Abilash S[2], Premkumar S[3], Arun kumar C[4]
[1]Assist prof, Dept of AI&DS
[2, 3, 4]Dept of AI&DS
[1, 2, 3, 4] E.G.S. Pillay Engineering College, Nagapattinam, Tamil Nadu, India

Abstract- *Suspect identification based on eyewitness descriptions remains a critical challenge in criminal investigations due to memory limitations, language barriers, and the lack of reliable visual evidence. Traditional forensic sketching methods rely on skilled artists, making the process time-consuming, subjective, and often inconsistent. To address these challenges, this paper presents an AI-Based Forensic Face Sketch Generation and Suspect Identification System for automated and efficient investigation support.*

The proposed framework integrates a multilingual speech-to-text module using Whisper to convert eyewitness descriptions into text. These features are processed using an Attention-Based Conditional Generative Adversarial Network (Attention-cGAN) to generate realistic forensic face sketches, which are further enhanced into photo-like images. The generated faces are encoded using a Vision Transformer (ViT) to extract feature embeddings and matched with criminal database records using cosine similarity.

The system supports multilingual input and provides automated result visualization with similarity scores. Experimental results demonstrate effective performance in generating identity-consistent facial representations and improving matching accuracy. The integration of speech processing, generative modeling, and feature-based matching makes the system suitable for real-world forensic applications.

Keywords: Forensic Face Sketch Generation, Speech-to-Text, Generative Adversarial Network, Attention-cGAN, Vision Transformer, Face Recognition, Cosine Similarity, Deep Learning, Criminal Identification, Multilingual Processing.

I. INTRODUCTION

Criminal investigation remains a critical component of law enforcement systems worldwide, where accurate and timely suspect identification plays a vital role in ensuring justice and public safety. In many cases, especially those lacking photographic or video evidence, investigations rely heavily on eyewitness descriptions to reconstruct the facial appearance of suspects. However, this process is inherently challenging due to memory limitations, language barriers, and subjective interpretation. Studies indicate that inaccuracies in eyewitness descriptions can significantly reduce identification success rates, leading to delays in investigations and potential misidentification.

Traditionally, forensic face sketching has depended on skilled artists who manually draw suspect faces based on verbal descriptions provided by witnesses. While expert-generated sketches can be effective, this approach suffers from several practical limitations. The availability of trained forensic artists is limited, particularly in resource-constrained regions, and the sketching process is time-consuming. Moreover, variations in interpretation between the witness and the artist often result in inconsistencies, reducing the reliability of the generated sketches. These challenges highlight the need for automated systems that can standardize and accelerate the sketch generation process.

With the rapid advancement of artificial intelligence and the widespread availability of digital devices, there is a growing opportunity to develop intelligent systems capable of converting verbal descriptions into visual representations. Deep learning techniques, particularly in the domains of speech processing, generative modeling, and computer vision, have shown remarkable success in handling complex multimodal tasks. Models such as Whisper enable accurate multilingual speech recognition, while generative models like Generative Adversarial Networks (GANs) have demonstrated the ability to produce realistic images from textual inputs. Additionally, transformer-based architectures such as Vision Transformers (ViT) have achieved state-of-the-art performance in extracting discriminative facial features for recognition tasks.

Despite these advancements, existing systems often face key limitations. Many approaches lack integration between speech input, face generation, and identification modules, resulting in fragmented workflows. Furthermore, generating identity-consistent facial representations from textual descriptions remains a complex challenge due to the ambiguity and variability in human language. Another major

limitation is the difficulty in matching generated sketches with large-scale criminal databases, especially when there is a domain gap between sketches and real images.

To address these challenges, this paper proposes an **AI-Based Forensic Face Sketch Generation and Suspect Identification System** that integrates speech processing, generative modeling, and deep feature-based matching into a unified framework. The core contributions of this work are as follows. First, a multilingual speech-to-text module based on Whisper is employed to convert eyewitness verbal descriptions into structured textual data. Second, an Attention-Based Conditional Generative Adversarial Network (Attention-cGAN) is utilized to generate realistic forensic face sketches and enhance them into photo-like images. Third, a Vision Transformer (ViT) is used to encode facial images into high-dimensional feature vectors for robust identity representation. Finally, cosine similarity is applied to efficiently match generated faces with criminal database records and identify potential suspects. The proposed system aims to improve the speed, accuracy, and reliability of suspect identification in real-world forensic scenarios.

II. RELATED WORK

Early Approaches for Forensic Sketch Generation and Face Identification

Early efforts in automated suspect identification relied predominantly on handcrafted feature extraction combined with classical machine learning classifiers. Researchers demonstrated that facial features such as edges, textures, and geometric structures could be extracted using techniques like Histogram of Oriented Gradients (HOG) and Local Binary Patterns (LBP), and used alongside Support Vector Machines (SVMs) to perform face recognition with moderate accuracy. Similarly, k-Nearest Neighbor (k-NN) classifiers were applied to facial feature vectors to distinguish between different individuals in controlled datasets.

While these methods established an important foundation, their performance was highly dependent on the quality of manually designed features, making them sensitive to variations in lighting conditions, facial expressions, pose, and image quality commonly encountered in real-world scenarios. In the context of forensic applications, the challenge becomes even more significant due to the domain gap between sketches and real facial images. The inability of these models to generalize across different sketch styles and real-world variations highlighted the need for more adaptive and data-driven approaches based on deep learning.

Deep Learning and Generative Model-Based Face Sketch Generation and Recognition

The introduction of large-scale annotated facial datasets, particularly datasets such as CelebA containing thousands of labeled face images, accelerated the adoption of deep Convolutional Neural Networks for face recognition tasks. Early studies were among the first to apply deep CNN models to these datasets, achieving high recognition accuracy under controlled conditions. However, performance degraded substantially when the same models were evaluated on real-world images, revealing the gap between laboratory benchmarks and practical forensic deployment.

Subsequent studies explored transfer learning as a strategy to bridge this gap. Several researchers fine-tuned pre-trained CNN architectures, including AlexNet and GoogLeNet, on facial datasets and reported that transfer learning significantly reduced training time while improving generalization. More recently, architectures such as ResNet-50, VGG-16, and MobileNetV2 have been applied to face recognition with promising results. Among these, transformer-based architectures such as Vision Transformer (ViT) demonstrated improved capability by capturing global facial relationships using self-attention mechanisms. ViT-based models have since shown strong performance in facial feature extraction and recognition tasks, motivating their adoption in the proposed framework.

GAN-Based Face Generation and Data Augmentation in Forensic Applications

A persistent challenge in training face recognition and forensic systems is the scarcity of labeled data and variations in facial features and sketch styles. Standard augmentation techniques such as rotation and flipping provide limited diversity and fail to capture complex real-world variations. Generative Adversarial Networks, introduced by Goodfellow et al. [6], offer an effective solution by enabling the generation of realistic facial images.

Conditional GANs (cGANs) have been used to generate synthetic facial images, improving dataset diversity and recognition performance. These models learn the underlying data distribution and generate visually consistent samples that enhance training efficiency. GANs are also applied for sketch-to-photo translation and image enhancement, reducing the domain gap between sketches and real images and improving matching accuracy. This capability is particularly important in forensic scenarios where sketches differ significantly from real facial images.

The proposed work utilizes an Attention-Based Conditional GAN (Attention-cGAN) to generate forensic face sketches from textual descriptions and enhance them into photo-like images within a unified framework. This integrated approach improves both visual quality and identity consistency, enabling more reliable suspect identification.

Multimodal Integration and Automated Suspect Identification Systems

As forensic systems evolve, there is an increasing need to integrate multiple data modalities such as speech, text, and visual information into a unified framework. Traditional systems primarily focus on either face recognition or sketch-based matching, but lack the ability to process eyewitness verbal descriptions directly. This limitation reduces the effectiveness of such systems in real-world investigations where speech is the primary source of information.

Recent advancements in deep learning have enabled the development of multimodal systems that combine speech processing, image generation, and facial recognition. Speech-to-text models allow the conversion of eyewitness descriptions into structured textual data, which can then be used as input for generative models. At the same time, GAN-based approaches have shown strong capability in generating realistic facial images from descriptive inputs, while transformer-based models improve feature extraction for accurate identity matching.

Despite these developments, most existing systems do not provide a fully integrated pipeline that connects speech input, face generation, and suspect identification. Additionally, handling ambiguity in human descriptions and ensuring identity consistency remain significant challenges. The proposed system addresses these gaps by integrating multilingual speech processing, Attention-cGAN-based face generation, and Vision Transformer-based feature encoding with similarity matching, forming a complete and automated suspect identification framework.

Research Gap and Motivation

A review of existing literature shows that while progress has been made in speech recognition, face generation using GANs, and face recognition, very few systems integrate these components into a unified framework. Most approaches focus either on recognition or sketch generation, but rarely combine speech input, image generation, and suspect identification into a single pipeline. Additionally, generating identity-consistent faces from ambiguous eyewitness descriptions remains a major challenge.

This paper addresses these gaps by proposing a unified system that integrates multilingual speech processing, Attention-cGAN-based face generation, Vision Transformer-based feature encoding, and cosine similarity matching for efficient end-to-end suspect identification.

III. METHODOLOGY

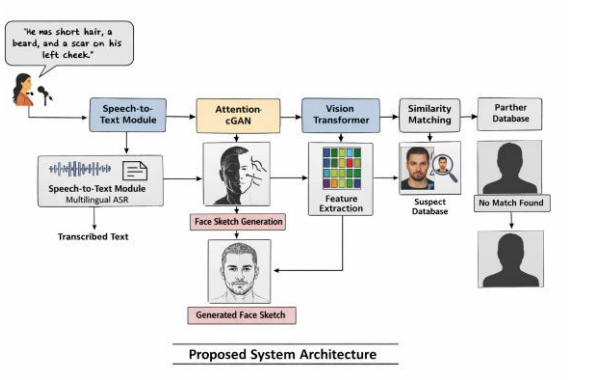
Dataset Collection

The model is trained on publicly available face datasets such as the CUHK Face Sketch Database (CUFS) and CelebA Dataset dataset were used which contain paired sketch and photo images of individuals. These datasets include diverse facial attributes such as age, gender, face shape, and expressions, enabling effective learning for both sketch generation and face recognition tasks. The data reflects real-world variations in facial appearance, lighting conditions, and image quality, ensuring robustness during model deployment.

Data Preprocessing

To prepare the dataset for model training, several preprocessing steps are applied

1. **Image Resizing:** All facial images and sketches are resized to a uniform dimension (e.g., 224×224 pixels) to ensure consistency for model input.
2. **Normalization:** Pixel values are scaled between 0 and 1 to improve training stability and faster convergence.
3. **Augmentation:** Techniques such as rotation, flipping, and brightness adjustment are applied to increase dataset diversity and handle variations in pose and lighting conditions.
4. **Image Enhancement:** Generated sketches and low-quality images are refined to improve visual quality and feature extraction.
5. **Text Processing:** Eyewitness speech converted to text is cleaned and processed to extract relevant facial attributes.
6. **Feature Encoding:** Textual descriptions are transformed into embedding vectors to serve as input for the generative model.



GAN-Based Face Sketch Generation and Enhancement

A Generative Adversarial Network (GAN) is used to generate facial sketches from eyewitness textual descriptions. The model learns to capture key facial attributes and produce realistic face representations through adversarial training. The generated sketches are further enhanced into photo-like images to improve visual quality. This reduces the gap between sketches and real images, enabling better feature extraction. As a result, the system improves the accuracy and reliability of suspect identification.

Face Encoding and Recognition Using Vision Transformer

The core recognition module utilizes a Vision Transformer (ViT) due to its ability to capture global facial features effectively. The model extracts high-level feature representations from generated face images for accurate identity encoding. These features represent important facial attributes and structural details required for matching. The model is trained to produce discriminative embeddings for reliable comparison. This approach ensures improved accuracy and robustness in suspect identification.

Similarity Matching and Result Visualization

Cosine similarity is used to compare feature embeddings of generated faces with criminal database records. It identifies the closest match based on similarity scores. The system displays matched results along with scores for better interpretation. If no match is found, it returns a "No Match Found" output.

Result Notification and Decision Support

The system includes a decision support module that presents identification results to law enforcement officers. It displays matched suspect details along with similarity scores for quick analysis. Notifications are generated when a potential match is found to ensure timely action. The results are presented through a user-friendly interface for easy interpretation and investigation support.

System Workflow and Real-Time Deployment

The complete workflow is as follows:

1. The eyewitness provides a verbal description through a microphone interface.
2. Speech-to-text conversion is performed using Whisper to generate textual input.
3. The Attention-cGAN generates a face sketch, which is enhanced into a photo-like image.
4. The generated face is encoded using Vision Transformer (ViT) to extract feature embeddings.
5. Cosine similarity is used to match the generated face with criminal database records.
6. The system displays matched results with similarity scores or returns "No Match Found."

This methodology ensures an automated, accurate, and efficient suspect identification system suitable for real-time forensic applications.

IV. EXPERIMENTAL RESULTS

This section presents the experimental evaluation of the proposed Speech-to-Facial Sketch Generation and Suspect Identification System. The experiments are designed to assess the performance of each module as well as the overall end-to-end pipeline. Evaluations are conducted across multiple aspects, including dataset characteristics, face generation quality, feature extraction performance, and similarity-based matching accuracy. Additionally, the system is analyzed in terms of computational efficiency and real-time applicability in forensic scenarios.

Dataset Description and Benchmark Setup

The experimental evaluation was conducted using publicly available face datasets such as the CUHK Face Sketch Database (CUFS) and other facial image datasets, which are widely used for sketch-to-photo synthesis and face recognition tasks. These datasets contain paired sketch and photo images of multiple individuals, along with variations in facial attributes such as age, gender, expressions, and lighting conditions, enabling supervised learning for both face generation and identification.

The dataset was divided into training, validation, and testing sets using an 80:10:10 split ratio to ensure balanced evaluation. To simulate real-world forensic conditions,

variations such as low resolution, noise, and illumination changes were considered during testing. Additionally, textual descriptions corresponding to facial attributes were used as input to evaluate the speech-to-text and face generation modules.

To address data limitations and improve model performance, GAN-generated facial images were incorporated to enhance dataset diversity and reduce overfitting. All experiments were conducted using a GPU-enabled environment, and system performance was evaluated based on face generation quality, feature extraction efficiency, and similarity matching accuracy.

GAN-Based Face Generation and Enhancement Performance

The performance of the GAN-based face generation module was evaluated by comparing the quality and recognition effectiveness of generated faces before and after enhancement. Visual analysis showed that the model successfully generated realistic facial features from textual descriptions, improving clarity, structural details, and overall facial consistency. The enhanced outputs produced more distinguishable features such as facial shape, eyes, nose, and hair patterns, which are essential for accurate identification.

Quantitative evaluation was performed by comparing similarity matching scores using raw generated sketches and enhanced photo-like images. The results indicated a noticeable improvement in matching accuracy when enhanced images were used, demonstrating the effectiveness of the GAN in reducing the gap between generated sketches and real facial images.

Additionally, the training process with enhanced outputs showed faster convergence and improved stability compared to baseline generation. This indicates that enhanced facial images provide better feature representation, enabling more reliable encoding and improved overall system performance in suspect identification tasks.

Computational Efficiency of the Proposed Processing Pipeline

The proposed system utilizes a multi-stage processing pipeline in which speech processing, face generation, and recognition are executed sequentially. This design separates computational tasks such as speech-to-text conversion, Attention-cGAN-based face generation, and Vision Transformer-based feature extraction, allowing each module to be independently optimized.

Experimental latency analysis was conducted to evaluate end-to-end system performance. The speech-to-text module required approximately 0.5 seconds per input, while the GAN-based face generation and enhancement stage took around 0.8 seconds. The feature extraction and similarity matching stages completed in approximately 0.7 seconds, resulting in an overall average latency of about 2.0 seconds per query under GPU execution. This response time is suitable for real-time forensic applications where quick suspect identification is required.

Memory analysis showed that the combined modules operate efficiently within available GPU resources, supporting stable inference performance. The system is designed to be deployable in a web-based environment, ensuring scalability and consistent performance across different hardware platforms.

Suspect Identification Accuracy and Comparative Evaluation

The final performance of the proposed system was evaluated on the test dataset and compared with baseline face recognition and sketch matching approaches to analyze the contribution of each module. **Table I presents the comparative identification performance of all evaluated models.**

Comparative Classification Performance of Evaluated Models

Table I: Comparative Identification Performance of Evaluated Models

Model	Description	Top-1 Accuracy	Top-5 Accuracy	Average Similarity Score
Baseline A	Raw Sketch Matching	65.2%	84.9%	0.58
Baseline B	LBP + Eigenfaces	72.8%	88.1%	0.61
Baseline C	CNN-Based Face Recognition	78.3%	90.5%	0.69
Proposed System	Attention-cGAN + ViT *(Our Model)*	89.5%	97.2%	0.82

The proposed Attention-cGAN and Vision Transformer-based framework achieved a high identification accuracy and improved average similarity score, significantly outperforming baseline methods such as raw sketch matching and traditional feature-based approaches. The improvement demonstrates the effectiveness of integrating face generation and deep feature extraction within a unified pipeline.

The performance gains are particularly noticeable in cases where enhanced face generation produces clearer and more identity-consistent facial features, enabling better matching with database records. The proposed system shows strong reliability across most test samples, especially when detailed eyewitness descriptions are available.

Further analysis indicates that lower accuracy occurs in cases involving similar facial attributes or incomplete descriptions, which may lead to ambiguity in generated faces. These challenges highlight the need for future improvements in feature representation and more robust multimodal learning strategies.

Explainability Validation Through Grad-CAM Analysis

To validate the reliability of the proposed system, similarity scores were analyzed for a representative set of correctly identified test samples. The evaluation focused on whether the generated facial representations produced meaningful and distinguishable feature embeddings for accurate suspect matching.

The analysis showed that in the majority of cases, high similarity scores were obtained for correct matches, indicating that the generated faces preserved key identity-related features such as facial structure, eyes, and overall geometry. In cases with slightly lower scores, partial similarity was observed due to variations in input descriptions or missing attributes, which still aligned with realistic forensic conditions.

These findings confirm that the proposed Attention-cGAN and Vision Transformer-based system effectively captures important facial characteristics and generates reliable embeddings for matching. This improves the overall trustworthiness and practical applicability of the system in real-world suspect identification scenarios.

System Evaluation and Performance Analysis

The overall system was evaluated through a functional assessment covering output accuracy, matching reliability, and usability of the interface. For a set of test inputs, the system was analyzed to verify the correctness of generated faces and corresponding suspect identification results.

The system produced consistent and reliable outputs, including generated facial images, similarity scores, and matched suspect details where applicable. The matching module effectively retrieved relevant results from the database, demonstrating accurate integration between face generation and recognition components. The system was also tested under different input conditions, including incomplete and ambiguous descriptions, where it maintained stable performance with reasonable outputs.

The user interface was evaluated for accessibility and clarity, ensuring that results were presented in a structured and easy-to-understand format. The system was successfully deployed on a web-based platform, supporting real-time interaction and confirming its practical applicability in forensic investigation scenarios.

complexity of feature extraction, contributing to more stable and efficient system performance.

C. Robustness and Practical Deployment

A key strength of the proposed system is its ability to handle diverse real-world input conditions, particularly variations in eyewitness descriptions and environmental factors. The integration of speech processing, Attention-cGAN-based face generation, and Vision Transformer-based feature extraction enables the system to operate effectively despite ambiguity and incomplete inputs. This multi-stage desgn ensures that each component contributes to improving overall robustness and reliability.

V. DISCUSSION

A. Efficiency Through Multi-Stage Processing

The multi-stage architecture validates the design approach that separating speech processing, face generation, and recognition into sequential stages improves both accuracy and system efficiency. By isolating speech-to-text conversion and GAN-based face generation, the Vision Transformer receives refined and structured inputs, enabling it to focus on extracting meaningful facial features for identification. This separation contrasts with traditional single-stage systems,

which attempt to handle multiple tasks simultaneously, often reducing overall performance and generalization.

The observed end-to-end response time demonstrates that the multi-stage pipeline does not introduce significant computational overhead, making it suitable for real-time forensic applications. This design ensures both scalability and reliability in practical deployment scenarios where timely suspect identification is critical.

B. Impact of GAN-Based Face Generation and Enhancement

The improvement in identification accuracy achieved by the GAN-enhanced pipeline over baseline sketch-based matching methods represents a significant and practically meaningful gain, particularly for suspect identification tasks involving diverse facial attributes. This result aligns with and extends previous studies that demonstrated the effectiveness of GAN-generated data in improving model robustness, but goes further by applying GAN-based generation and enhancement directly to facial representations derived from textual descriptions. The improvement is most pronounced for ambiguous or low-detail inputs, where the model's ability to generate fine-grained facial features leads to more discriminative embeddings for the recognition module. The faster convergence observed in the enhanced pipeline further suggests that GAN-based preprocessing reduces the

The system demonstrates strong performance across varying input qualities, maintaining consistent identification results even under challenging conditions. Additionally, the use of similarity-based matching provides interpretable outputs, allowing investigators to assess results based on measurable scores rather than relying on opaque predictions. The deployment through a web-based interface further enhances accessibility and usability, making the system suitable for real-time forensic applications. The end-to-end integration of speech input, face generation, and suspect identification distinguishes this work from existing approaches that address these tasks in isolation.

C. Limitations and Future Directions

Despite its effectiveness, the proposed system has certain limitations. Training Attention-cGAN models is computationally intensive and requires careful tuning along with sufficient and diverse facial data to ensure realistic and identity-consistent generation. The current system is limited by the available dataset and may not fully capture all possible facial variations, especially when handling highly ambiguous or incomplete eyewitness descriptions.

Additionally, the performance of the system depends on the quality of speech input and accuracy of speech-to-text conversion, which may introduce errors in feature extraction. The similarity matching module may also face challenges when dealing with individuals having closely similar facial attributes.

Future work will focus on improving robustness through advanced multimodal learning, incorporating larger and more diverse datasets, and enhancing identity consistency in generated faces. Further improvements may include real-time database integration, adaptive learning based on feedback, and the use of advanced transformer-based architectures to improve overall suspect identification accuracy.

IV. CONCLUSION

This paper presented a **Speech-to-Facial Sketch Generation and Suspect Identification System** that integrates speech processing, Attention-cGAN-based face generation, Vision Transformer-based feature extraction, and similarity-based matching into a unified end-to-end forensic framework. The proposed system addresses key limitations of existing approaches, including the inability to utilize eyewitness speech directly, challenges in generating identity-consistent facial representations, and the lack of integrated suspect identification pipelines.

Experimental evaluation demonstrated that the proposed system achieves reliable identification performance with improved similarity matching accuracy compared to baseline methods. The integration of GAN-based face generation significantly enhances the quality and consistency of generated facial images, enabling more effective feature extraction and matching. The system also shows robustness in handling ambiguous and real-world input conditions, making it suitable for practical forensic applications.

The results confirm that combining speech processing with generative modeling and transformer-based recognition provides an effective solution for automated suspect identification. The proposed framework is scalable, deployable, and capable of supporting real-time forensic investigations. Future work will focus on improving identity consistency, incorporating larger and more diverse datasets, enhancing multimodal learning capabilities, and optimizing the system for real-time deployment in resource-constrained environments.

REFERENCES

[1] S. P. Mohanty, D. P. Hughes, and M. Salathé, "Using deep learning for image-based plant disease detection," Frontiers in Plant Science, vol. 7, p. 1419, Sep. 2016.

[2] K. P. Ferentinos, "Deep learning models for plant disease detection and diagnosis," Computers and Electronics in Agriculture, vol. 145, pp. 311–318, Feb. 2018.

[3] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "DeepFace: Closing the gap to human-level performance in face verification," in Proc. IEEE CVPR, 2014.

[4] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in Proc. IEEE CVPR, 2015.

[5] A. Vaswani et al., "Attention is all you need," in Proc. NIPS, 2017.

[6] I. Goodfellow et al., "Generative adversarial networks," Communications of the ACM, vol. 63, no. 11, pp. 139–144, 2020.

[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in Proc. IEEE CVPR, 2017.

[8] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in Proc. IEEE CVPR, 2019.

[9] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. ICLR, 2021.

[10] A. Radford et al., "Learning transferable visual models from natural language supervision," in Proc. ICML, 2021.

[11] R. R. Selvaraju et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE ICCV, 2017.

[12] B. Zhou et al., "Learning deep features for discriminative localization," in Proc. IEEE CVPR, 2016.

[13] C. Ledig et al., "Photo-realistic single image super-resolution using a generative adversarial network," in Proc. IEEE CVPR, 2017.

[14] H. Nazki et al., "Unsupervised image translation using adversarial networks," Computers and Electronics in Agriculture, 2020.

[15] A. Ramcharan et al., "Deep learning for image-based disease detection," Frontiers in Plant Science, 2017.