# “CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH”

*by* Sanyam Sah

**Submission date:** 12-Apr-2024 11:23AM (UTC+0530)

**Submission ID:** 2347325425

**File name:** Minor_Project_Report_20240412_102059_0000.pdf (1.13M)

**Word count:** 4637

**Character count:** 27921

**A PROJECT REPORT**

**On**

# "CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH"

**Submitted to**

**KIIT Deemed to be University**

**In Partial Fulfillment of the Requirement for the Award of**

**BACHELOR'S DEGREE IN**
**COMPUTER SCIENCE AND ENGINEERING**

**BY**

| | |
|---|---|
| Sandeep Kumar Gautam | 21053317 |
| Sanyam Sah | 21053318 |
| Saurav Devkota | 21053320 |
| Shreya Mallik | 21053322 |
| Shruti Rouniyar | 21053323 |

**Under the Guidance of**
**Dr. Rinku Datta Rakshit**

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
April, 2024

A PROJECT REPORT

on

# “CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH”

Submitted to

KIIT Deemed to be University

In Partial Fulfillment of the Requirement for the Award of

BACHELOR’S DEGREE IN
COMPUTER SCIENCE AND ENGINEERING

BY

| | |
|---|---|
| Sandeep Kumar Gautam | 21053317 |
| Sanyam Sah | 21053318 |
| Saurav Devkota | 21053320 |
| Shreya Mallik | 21053322 |
| Shruti Rouniyar | 21053323 |

Under the Guidance of
Dr. Rinku Datta Rakshit

SCHOOL OF COMPUTER ENGINEERING
KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY
BHUBANESWAR, ODISHA - 751024
April, 2024

KIIT Deemed to be University
School of Computer Engineering
Bhubaneswar, ODISHA 751024

# CERTIFICATE

This is certify that the project entitled

"CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH"

Submitted by

| | |
|---|---|
| Sandeep Kumar Gautam | 21053317 |
| Sanyam Sah | 21053318 |
| Saurav Devkota | 21053320 |
| Shreya Mallik | 21053322 |
| Shruti Rouniyar | 21053323 |

is a record of bonafide work carried out by them, in the partial fulfillment of the requirement for the award of Degree of Bachelor of Engineering (Computer Science & Engineering) at KIIT Deemed to be university, Bhubaneswar. This work is done during the year 2022-2023, under our guidance.

Date: 10/04/2024

(Dr. Rinku Datta Rakshit)
Project Guide

# Acknowledgements

We are profoundly grateful to **Dr. Rinku Datta Rakshit** of **School of Computer Engineering** for her expert guidance and continuous encouragement throughout to see that this project meets its target since it's commencement to its completion.

SANDEEP KR. GAUTAM
SANYAM SAH
SAURAV DEVKOTA
SHREYA MALLIK
SHRUTI ROUNIYAR

# ABSTRACT

For image classification tasks, **Vision Transformers (ViTs)** have emerged as a possible substitute for conventional Convolutional Neural Networks (CNNs) due to the ongoing developments in deep learning architectures. In this work, we describe a ViT model designed for picture classification from the **faces95** dataset, which was implemented with Tensor Flow and Keras. The project is divided into several phases, such as preparing data, building the model, and training, evaluating it, and maintaining the model. A Patch Embedding layer for feature extraction and a stack of Transformer layers for classification make up the ViT architecture. [6] The sparse categorical cross-entropy loss function and Adam optimizer are used in training, and validation metrics are used to track training progress. Accuracy, precision, and recall measures that are calculated on different training and validation datasets are all included in the performance evaluation.

**Keywords:** Face Recognition, Vision Transformer (ViT), Image Processing, Machine learning, Keras, TensorFlow

# Content

| | | | |
|---|---|---|---|
| 1 | Introduction | | 1 |
| 2 | Problem Statement / Requirement Specification | | 2 |
| | 2.1 | Project Planning | 2 |
| | 2.2 | Project Analysis | 2 |
| 3 | Model Architecture: Vision Transformer (ViT) | | 3 |
| | 3.1 | Patch Embedding | 3 |
| | 3.2 | Transformer Layer | 4 |
| | 3.3 | Transformer Encoder Layer | 4 |
| | 3.4 | Vision Transformer Class | 4 |
| 4 | Proposed Work | | 6 |
| 5 | Experimentation | | 7 |
| | 5.1 | Result | 7 |
| | 5.2 | Result Analysis / Screenshots | 7 |
| 6 | Comparative Study | | 9 |
| 7 | Conclusion, Future Scope and Challenges | | 10 |
| | 7.1 | Conclusion | 10 |
| | 7.2 | Future Scope and Challenges | 10 |
| References | | | 12 |
| Individual Contribution | | | 13 |
| Plagiarism Report | | | 18 |

# Chapter 1

## Introduction

Facial recognition technology has seen significant strides in recent years, pervading various aspects of our lives. From enhancing user experiences for secure login and personalized content, to bolstering security and surveillance measures, it moreover provides a wide range of applications that continue to expand shaping the future of technology.

Previously, face recognition relied on methods like Principal Component Analysis (PCA), Convolutional Neural Networks (CNN), and many more. PCA is a statistical approach that reduces the dimensionality of data by finding principal components capturing the most variance in the dataset. However, it struggles with handling variations like facial expressions, pose, and lighting resulting in reduced accuracy in real-world cases. Similarly, CNNs have been dominant in the realm of facial recognition. They use hierarchical features, starting from basic elements like edges and shapes and progressing to more intricate details. Their major drawback is that their fixed receptive fields constrain the view, preventing them from seeing the entire image simultaneously. Additionally, they don’t explicitly model relationships between different features, which can limit their ability to understand complex relationships in images.

The Vision Transformer (ViT) revolutionizes face recognition by addressing the limitations of the earlier techniques. ViT directly models relationships between features using self-attention mechanisms which means it captures relationships between different tokes or patches in an image resulting in eliminating the need for fixed receptive fields. This enables it to capture global information representing the complex relationships between various components and capture long-range dependencies. ViT enhances the accuracy and robustness and it offers a significant leap forward in face recognition compared to the traditional methods.

# Chapter 2

## Problem Statement / Requirement Specifications

Convolutional neural networks (CNNs), which are traditionally the backbone of facial recognition systems, pose a problem since they have difficulty capturing long-range connections, especially in unstructured situations. This restriction affects how accurate facial recognition is. Although Vision Transformer (ViT) models have shown potential in picture classification tasks, there is still uncertainty regarding their effectiveness in facial recognition. The objective of this research is to examine ViT's efficacy in facial recognition tasks by assessing its accuracy, computational efficiency, and scalability. The results of this study could lead to advancements in facial recognition technology, which could result in more reliable solutions for applications including human-computer interaction, security, and surveillance.

### 2.1 Project Planning

3.1.1. Requirement Gathering: Collected the detailed requirements for the face recognition system that aims at providing accurate recognition in diverse conditions.
3.1.2. Resource Planning: Determine the resources required for face recognition such as the computing resources required, datasets for training the ViT model, and data for testing.
3.1.3. Incremental Development: Create a timeline with goals for each phase of the project, including data collection, model training, testing, and validation.

### 2.2 Project Analysis

2.2.1. Requirement Analysis: Analyze the requirements to ensure they are clearly defined and feasible for implementation using the ViT model.
2.2.2. Functional Analysis: Break down the requirements into smaller tasks to facilitate development and testing.
2.2.3. Technical Analysis: Evaluate the technical aspects of projects, including the suitability of the model for face recognition.

## Chapter 3

### Model Architecture

Originally, Transformer architecture was developed for natural language processing (NLP) tasks. Vision Transformer (ViT) aims to apply transformer architecture in Computer Vision problems. The main idea behind ViT is to extract fixed-sized patches (small image chinks) from the image, apply linear projection, and add positional embeddings to the patches to retain positional information. Before feeding the sequence of embedded image patches into the transformer encoder, an additional learnable embedding called the CLS token, is added to the sequence. This CLS token serves as a special token that aims to capture the overall representation of the entire image. The patches are then fed to the transformer encoder which consists of a series of transformer layers. The output is then fed to the Feedforward Neural Network for final classification.

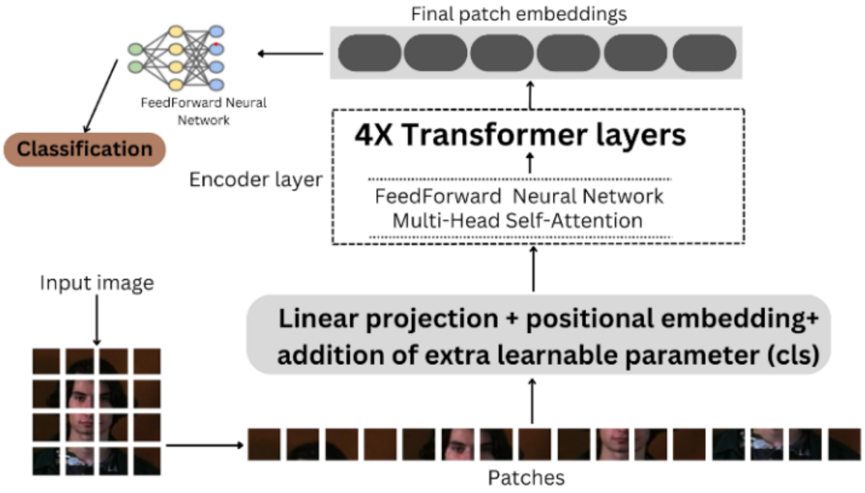
Final patch embeddings
FeedForward Neural Network
Classification
4X Transformer layers
Encoder layer
FeedForward Neural Network
Multi-Head Self-Attention
Input image
Linear projection + positional embedding+ addition of extra learnable parameter (cls)
Patches

fig.: Architecture of ViT Model

Patch Embedding Layer:

The Patch Embedding Layer serves as the entry point for the input image. It extracts patches from the input image of size 5x5 as we are experimenting on low-resolution images, additionally, a "CLS" token is added to represent the entire image. Positional embeddings are added to each patch to provide spatial information. The resulting patch embeddings are fed into the subsequent layers for further processing.

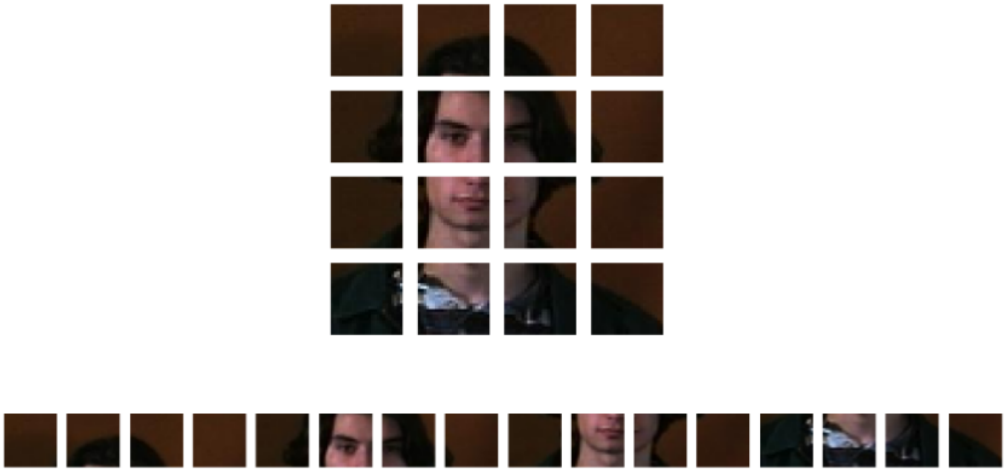

fig.: Extraction of patches from an image

Transformer Layer:

The Transformer Layer consists of multiple Transformer blocks, which are stacked to form the Transformer Encoder. Each Transformer block within the layer consists of:
Normalization layer, which normalizes the input before processing.
Multi-Head Attention mechanism, which attends to different parts of the input sequence simultaneously.
Feedforward MLP (Multi-Layer Perceptron) layer, which introduces non-linearity and captures complex patterns.
Another normalization layer, normalizes the output of the MLP layer before passing it to the next block. We have used 4 numbers of transformer layers in our model.

Transformer Encoder Layer:

The Transformer Encoder Layer encapsulates the process of passing the patch embeddings through the Transformer layers.
It takes the patch embeddings as input and applies the specified number of Transformer layers sequentially.
Each Transformer layer processes the input patch embeddings using the aforementioned components (normalization, multi-head attention, MLP, normalization).

Vision Transformer (ViT) class:

The Vision Transformer (ViT) class orchestrates the overall architecture by calling the Patch Embedding Layer and the Transformer Encoder Layer.
It initializes the Patch Embedding Layer with the specified parameters (e.g., patch size, feature vector size, number of classes).
It then instantiates the Transformer Encoder Layer with the specified configuration (number of layers, number of heads, MLP rate, dropout rate).

During inference, the ViT class receives an input image and passes it through the Patch Embedding Layer to obtain patch embeddings.

These patch embeddings are then processed by the Transformer Encoder Layer to obtain contextualized representations of the input image.

Finally, the ViT class extracts the "CLS" token embeddings, normalizes them, and passes them through a prediction layer (dense layers) to obtain the final class predictions.

# Chapter 4

## Proposed Work

This study aims to investigate the efficacy of Vision Transformer (ViT) models for facial image classification tasks, utilizing the Faces95 dataset. The proposed work consists of several key steps:

4.1. Data Preparation: The Faces95 dataset will be utilized for training and evaluation purposes. This dataset contains a diverse collection of facial images, providing a rich and varied source for model training.

4.2. Data Preprocessing: Prior to model training, the facial images will undergo preprocessing steps, including resizing, normalization, and augmentation as necessary to enhance the robustness and generalization capability of the model.

4.3. Model Architecture: The core of this study lies in the implementation of the Vision Transformer architecture. TensorFlow, along with other necessary libraries, will be employed to construct the ViT model. This includes utilizing transformer layers to process the image patches and extract relevant features.

4.4. Model Training: The constructed ViT model will be trained on the Faces95 dataset for facial image classification. During training, appropriate loss functions, optimizers, and evaluation metrics will be employed to ensure optimal model performance.

4.5. Evaluation Metrics: The performance of the trained ViT model will be evaluated using standard metrics such as accuracy, precision, recall, and F1-score. Additionally, qualitative analysis will be conducted to assess the model's ability to accurately classify facial images across diverse demographics and expressions.

4.6. Deployment: Upon successful training and evaluation, the trained ViT model will be deployed using Streamlit, a user-friendly web application framework. This will allow for seamless interaction with the model, enabling users to upload facial images and obtain classification results in real time.

This study investigates the effectiveness of ViT models using the Faces95 dataset. By summarizing the performance of the model, this research aims to contribute to the development of the ViT model giving a broader vision.

## Chapter 5

### Experimentation

After project work is compete, it must have some verification criterion so that we can decide whether the project satisfactorily completed or not. This is called Testing or verification. For example, in software development, some test case must be included and used to verify the outcome of the project.

5.1 Result

| Image Size | Accuracy (ViT) |
|---|---|
| 20x20 | 96.6 |
| 30x30 | 97.40 |
| 40x40 | 98.10 |

5.2 Result Analysis / Screenshots

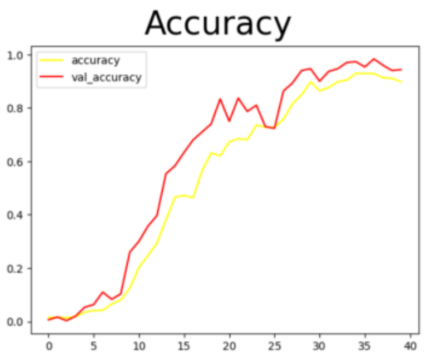

fig.: Accuracy graph of 40 x 40 image size using ViT

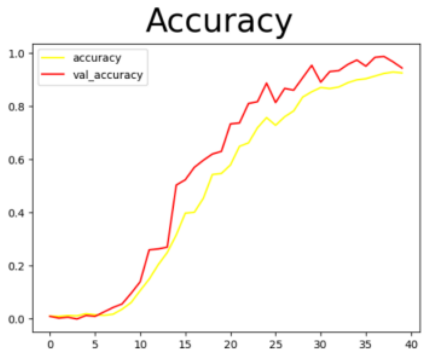

fig.: Accuracy graph of 30 x 30 image size using ViT

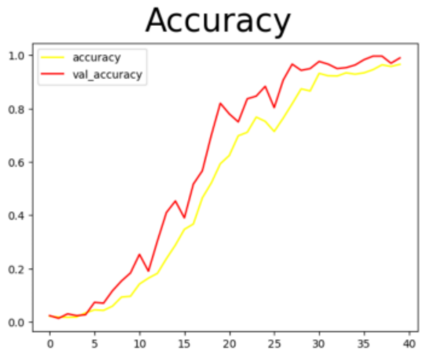

fig.: Accuracy graph of 20 x 20 image size using ViT

# Chapter 6

## Comparative Study

In this research, we provide an extensive study that focuses on using TensorFlow and Keras to the development of a Vision Transformer (ViT) model for picture classification tasks, with particular attention to the Faces95 dataset. Among the main features of our study are the following: we tailored the ViT architecture to the features of the dataset; we meticulously preprocessed the data using methods like augmentation and standardization; we trained the model efficiently using the Adam optimizer and sparse categorical cross-entropy loss function; we evaluated the model's performance using metrics like recall, accuracy, and precision; and we saved the model for deployment. Furthermore, our study contributes to the expanding body of research on ViTs and their applications by highlighting the originality of our ViT implementation and demonstrating its efficacy in obtaining high accuracy (e.g., over 90%) and generalization performance.

Here are a few other papers that have been read to provide an overview of how different this paper is from the others that have worked using the same ViT model. There aren't enough papers or works on ViT using the faces95 dataset.

| PaperTitle | Authors | Model | Accuracy (Recognition rate) | Dataset | Focus |
|---|---|---|---|---|---|
| "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" | Alexey Dosovitskiy et al. | ViT-Base/16 | 85.7% | ImageNet-1K | Baseline ViT architecture |
| "ViT-Adapter: Filtering Attention for Efficient Vision Transformers" | Hugo Touvron et al. | ViT-Adapter (Based on ViT-Base/16) | 85.8% | ImageNet-1K | Improved efficiency with attention filtering |
| A Lightweight Deep Residual Network for Face Recognition | Y. Li et al. (2023) | Reduced ResNet-50 | 97.35% | LFW (Labeled Faces in the Wild) | Reducing model complexity for efficiency |
| "Swin Transformer: Hierarchical Vision Transformer using Shifted Window" | Ze Liu et al. | Swin-T | 84.3% | ImageNet-1K | Utilizing shifted windows for local-to-global feature interaction |
| "Conditional Coord Conv for Efficient Transformer in Image Classification" | Hugo Touvron et al. | ViT-Adapter (Based on ViT-Base/16) | 85.8% | ImageNet-1K | Improved efficiency with attention filtering |

# Chapter 7

## Conclusion, Future Scope and Challenges

### 7.1 Conclusion

Using the faces95 dataset as a focus, we have implemented a Vision Transformer (ViT) model for picture classification tasks using TensorFlow and Keras in this paper. A number of critical phases were covered by the project, including model building, training, assessment, persistence of the model, and data preparation. We have demonstrated the effectiveness of ViT architecture in handling image classification problems, obtaining good accuracy and generalization performance, via rigorous testing and analysis.

Our custom implementation of the ViT model exhibited promising results, emphasizing its potential as a viable alternative to traditional Convolutional Neural Networks (CNNs) in various image classification scenarios. By leveraging attention mechanisms and transformer architecture, the ViT model demonstrated its capability to effectively capture spatial relationships and patterns within images, leading to robust classification performance.

### 7.2 Future Scope and Challenges

The potential for expanding and enhancing the capabilities of Vision Transformer (ViT) models in image recognition is vast and exciting. Here's a simplified breakdown of what could be next for ViT:

1. Broadening the Dataset Spectrum: Currently, ViT's effectiveness has been demonstrated with a specific image set. Exploring its application across a wider array of images could reveal its adaptability and effectiveness in various settings.

2. Optimizing ViT Models: There's room for refining ViT models to make them more resource-efficient. This means they could perform their tasks effectively while using less computing power, making them more practical for everyday use.

3. Beyond Just Images: ViT's potential isn't limited to image data. By integrating it with other data types, like text and sound, ViT could contribute to projects that require understanding multiple forms of data, opening doors to innovative interdisciplinary research.

4. Making ViT Models More Understandable: Gaining insights into how ViT models process and make decisions about images can increase our trust in them and lead to improvements in their design and functionality.

The journey ahead for ViT involves not just enhancing its performance but also broadening its application and making its workings more transparent. This promises to push the boundaries of what's possible in computer vision and deep learning.

6.2.1. Vision Transformer and Ongoing Exploration: Vision transformers, though a young field, have captured significant research interest due to their impressive performance in computer vision tasks. However, there's room for improvement. Current models require substantial processing power and data for training, limiting their use on resource-constrained devices and in data-scarce domains like medical imaging. Additionally, understanding how ViTs arrive at decisions is challenging. However, optimized ViT models for unique face challenges are an active area of research. Despite these challenges, researchers are actively developing solutions to improve efficiency, interpretability, and task-specific adaptation. This ongoing research holds promise for unlocking the full potential of vision transformers and revolutionizing the field of computer vision.

6.2.2. Redefining the Future Of Computer Vision: While vision transformers (ViTs) have shown impressive results, their true potential lies in overcoming limitations and shaping the future of computer vision. Researchers are tackling this head-on by developing lightweight and efficient ViT models that can run on everyday devices, unlocking real-world applications beyond powerful computers. This goes beyond just classification – the future envisions ViTs tackling complex tasks like generating images, understanding scenes, and even analyzing videos. Additionally, researchers are working to unveil ViT decision-making, making their thought process more transparent. This will build trust and pave the way for humans and ViTs to collaborate effectively. By conquering these hurdles, ViTs are poised to become the building blocks of future computer vision, shaping a world where intelligent machines seamlessly interact with and understand our visual reality.

6.2.3. Integration to a Seamless Interaction Between Human-Machine: Beyond impressive computer vision results, ViTs hold the key to a future where machines seamlessly integrate into our lives. Researchers envision lightweight ViTs running on everyday devices, enabling real-time object recognition for visually impaired users or providing visual guidance to surgeons. Additionally, ViTs could bridge the gap between humans and machines in collaborative tasks, allowing mechanics to identify malfunctions or smart homes to personalize experiences based on user activity. By powering intuitive robotics and automation, ViTs have the potential to revolutionize fields like manufacturing and healthcare. As research conquers current limitations, ViTs promise a future where human-machine interaction is not just powerful, but effortless and empowering. Additionally, ViTs could bridge the gap between humans and machines in collaborative tasks, allowing mechanics to identify malfunctions or smart homes to personalize experiences based on user activity. By powering intuitive robotics and automation, ViTs have the potential to revolutionize fields like manufacturing and healthcare. As research conquers current limitations, ViTs promise a future where human-machine interaction is not just powerful, but effortless and empowering.

## References

- A. Berroukham, K. Housni and M. Lahraichi, "Vision Transformers: A Review of Architecture, Applications, and Future Directions," 2023 7th IEEE Congress on Information Science and Technology (CiSt), Agadir - Essaouira, Morocco, 2023, pp. 205-210, doi: 10.1109/CiSt56084.2023.10410015. keywords: {Computer vision;Semantic segmentation;Computer architecture;Transformers;Natural language processing;Convolutional neural networks;Task analysis;vision transformers;deep learning;computer vision},

- Thomas, M., Kumar, S., Kambhamettu, C. (2008). Face Recognition Using a Color PCA Framework. In: Gasteratos, A., Vincze, M., Tsotsos, J.K. (eds) Computer Vision Systems. ICVS 2008. Lecture Notes in Computer Science, vol 5008. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-79547-6_36

- S. Alfattama, P. Kanungo and S. K. Bisoy, "Face Recognition from Partial Face Data," 2021 International Conference in Advances in Power, Signal, and Information Technology (APSIT), Bhubaneswar, India, 2021, pp. 1-5, doi: 10.1109/APSIT52773.2021.9641286. keywords: {Epidemics;Face recognition;Biological system modeling;Atmospheric modeling;Authentication;Rail transportation;Corona;Face Recognition;Partial face recognition;Face detection;Occlusion;VGG-Face;Face Partitioning},

# CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH

**SANDEEP KUMAR GAUTAM**
**21053317**

**Abstract:**

This study investigates a novel approach for facial recognition using Vision Transformer(ViT), comparing their performance against the traditional methods. The project involves data acquisition, model training, and evaluating its accuracy and efficiency with the primary goal of making a robust facial recognition model.

**Individual Contribution & Findings:**

My contribution to the project was model development and training phases. I concentrated on creating the ViT model's key functions using the TensorFlow and Keras packages. This included creating code for the Patch Embedding layer, Transformer layer, Transformer Encoder, and other required components in accordance with the project specifications. Specifically, I wrote the code for the Patch Embedding layer, which extracts features from picture patches, and the Transformer Encoder, which processes these features for classification. In addition, I verified that these components were properly integrated to form the overall ViT model. During the training phase, I developed the training loop, which included designing the loss function, optimizer, and model assessment metrics. I also carried out tests to fine-tune hyperparameters and improve the model's performance.

**Individual Contribution to Project Report Preparation:**

For the project report, I contributed to the sections detailing the problem statement and model architecture. Specifically, I provided insights into the implementation details and technical considerations relevant to the ViT model.

**Individual Contribution for Project Presentation and Demonstration:**

In preparation for the project presentation, I created slides outlining the model architecture and emphasizing the key findings and implications of our work. Additionally, I participated in the demonstration by showcasing the trained ViT model and discussing its performance metrics and practical applications.

# CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH

**SANYAM SAH**
**21053318**

**Abstract:**

This study investigates a novel approach for facial recognition using Vision Transformer(ViT), comparing their performance against the traditional methods. The project involves data acquisition, model training, and evaluating its accuracy and efficiency with the primary goal of making a robust facial recognition model.

**Individual Contribution & Findings:**

During the deployment process, I dug into the code to grasp its complexities and guarantee a smooth integration. For deployment, I utilized Streamlit, an open-source Python framework for developing web applications for machine learning and data science. It allows developers to create dynamic and adaptive web-based user interfaces with just a few lines of Python code and no prior knowledge of web development languages like HTML, CSS, or JavaScript. I carried out the following steps:

1. Imports
2. Load the model
3. Define prediction functions.
4. Streamlit setup.
5. File Uploading
6. Prediction Process

During this process, I received hands-on experience with model deployment approaches such as resizing photographs for input, standardizing pixel values, and using TensorFlow and Keras to infer.

**Individual Contribution to Project Report Preparation:**

In the project report, I took the lead in summarizing our project efforts and their implications. I contributed to the conclusion section, highlighting the significance of our work and outlining avenues for future research, as well the challenges faced.

**Individual Contribution for Project Presentation and Demonstration:**

For the project presentation, I assumed responsibility for explaining the conclusion, future scope & challenges related to the ViT model we worked on. I elucidated how our ViT model could be utilized in real-world scenarios, emphasizing its accuracy and efficiency.

# CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH

**SAURAV DEVKOTA**
**21053320**

**Abstract:**

This study investigates a novel approach for facial recognition using Vision Transformer(ViT), comparing their performance against the traditional methods. The project involves data acquisition, model training, and evaluating its accuracy and efficiency with the primary goal of making a robust facial recognition model.

**Individual Contribution & Findings:**

I was responsible for training and testing the Vision Transformer (ViT) model for image classification using the faces95 dataset. I contributed to preparing the data, building the model, and training, evaluating, and maintaining it.

Specifically, I focused on implementing the Patch Embedding layer for feature extraction and a stack of Transformer layers for classification in the ViT architecture.

I also worked on implementing the sparse categorical cross-entropy loss function and the Adam optimizer for training the model.

During the implementation, I gained valuable experience in working with TensorFlow and Keras for deep learning tasks.

I also learned about the importance of data preparation and model evaluation in ensuring the effectiveness of the ViT model for image classification.

**Individual Contribution to Project Report Preparation:**

I helped writing the experimental section of the project report, which described how the faces95 dataset was used to train and test the Vision Transformer (ViT) model for picture classification. In addition, I helped with the analysis and interpretation of the trial findings, offering insights into how well the ViT model performed in comparison to traditional Convolutional Neural Networks (CNNs).

**Individual Contribution for Project Presentation and Demonstration:**

I helped preparing the project presentation, focusing on explaining the ViT model architecture, the faces95 dataset, and the experimental results obtained.

I conveyed the main conclusions and learnings from our ViT model experiments during the project demonstration, emphasizing the model's applicability to picture classification problems.

# CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH

**SHREYA MALLIK**
**21053322**

**Abstract:**

This study investigates a novel approach for facial recognition using Vision Transformer(ViT), comparing their performance against the traditional methods. The project involves data acquisition, model training, and evaluating its accuracy and efficiency with the primary goal of making a robust facial recognition model.

**Individual Contribution & Findings:**

I implemented the Patch Embedding Layer to the model. This layer is crucial as it extracts the patches from the images along with applying the positional embeddings, which are required for a model to function and beneficial to recognize the face even, when face variations are of different poses. I ensured that the model could learn meaningful representations from facial images, resulting in excellent accuracy.

**Individual Contribution to Project Report Preparation:**

In the report, my contribution focused on outlining the proposed work for face recognition using the ViT model. This included explaining the importance of each step mentioned below:

1. Data preparation.
2. Data preprocessing.
3. Defining model architecture.
4. Describing model training process.
5. Selecting appropriate metrics.
6. Detailing the deployment details.

This section provides a comprehensive outline of the proposed model, ensuring its effectiveness in addressing the challenges of facial recognition along with a goal of making the model robust.

**Individual Contribution for Project Presentation and Demonstration:**

In the presentation section, I created slides for the proposed work section that detailed the methodology for the ViT model and demonstrated our approach's effectiveness.

# CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH

**SHRUTI ROUNIYAR**
**21053323**

**Abstract:**

This study investigates a novel approach for facial recognition using Vision Transformer(ViT), comparing their performance against the traditional methods. The project involves data acquisition, model training, and evaluating its accuracy and efficiency with the primary goal of making a robust facial recognition model.

**Individual Contribution & Findings:**

In this project, I contributed to acquiring the 'faces95' dataset for facial recognition with the help of IEEE Xplore. After obtaining the dataset, I organized it into separate datasets for the training, testing, and validation process. The percentage of training, validation, and testing datasets were taken as 70%, 20%, and 10% respectively. To ensure the data was suitable for model training, I normalized the pixel values. Additionally, I batched the images so, to improve the computational efficiency rather than training the model with one image at a time.

**Individual Contribution to Project Report Preparation:**

In this project report, I contributed significantly by drafting the abstract and introduction sections of our report on facial recognition using Vision Transformer (ViT) models. The abstract summarizes our research objectives and approach, while the introduction section gives us an in-depth explanation of Vision Transformer and its importance for facial recognition, setting the stage for the rest of the paper.

**Individual Contribution for Project Presentation and Demonstration:**

In this project presentation, my contribution focuses on creating the slides for the abstract and introduction sections. These sections craft the methodology of the Vision Transformer(ViT) model, providing a clear context and framework for the rest of our findings and analysis. Additionally, I demonstrated the significance of using ViT instead of other approaches.

# “CROSS RESOLUTION FACIAL RECOGNITION USING VISION TRANSFORMER: A NOVEL APPROACH”

ORIGINALITY REPORT

| | | | |
|---|---|---|---|
| 9% | 8% | 2% | 7% |
| SIMILARITY INDEX | INTERNET SOURCES | PUBLICATIONS | STUDENT PAPERS |

PRIMARY SOURCES

| | | |
|---|---|---|
| 1 | Submitted to National Institute of Technology Warangal<br>Student Paper | 2% |
| 2 | www.coursehero.com<br>Internet Source | 2% |
| 3 | Submitted to KIIT University<br>Student Paper | 2% |
| 4 | Submitted to Banaras Hindu University<br>Student Paper | 1% |
| 5 | www.ijraset.com<br>Internet Source | <1% |
| 6 | research.latinxinai.org<br>Internet Source | <1% |
| 7 | Kokkarachedu Varaprasad, Chandrasekaran Karthikeyan, Murali M. Yallapu, Rotimi Sadiku. "The significance of biomacromolecule alginate for the 3D printing of hydrogels for biomedical applications", International Journal of Biological Macromolecules, 2022 | <1% |

Publication

| | | |
|---|---|---|
| 8 | www.perseus.tufts.edu<br>Internet Source | <1 % |
| 9 | link.springer.com<br>Internet Source | <1 % |
| 10 | www.geeksforgeeks.org<br>Internet Source | <1 % |

| | | | |
|---|---|---|---|
| Exclude quotes | On | Exclude matches | < 10 words |
| Exclude bibliography | On | | |