Major Project Report

On

DEEP LEARNING BASED FACE EXTRACTION AND ITS ENHANCEMENT FOR VIDEO SURVEILLANCE

Submitted in the partial fulfillment for the award of degree of Bachelor of Technology

In

Electronics and Communication Engineering By

Tejas Bibekar (20106010)
Tanmay Giram (20106021)
Nishant Wankhade (20106070)

B. Tech, VIII Semester

Under the guidance of

Jitendra Bharadwaj
(Asst. Professor)

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING SCHOOL OF STUDIES

IN ENGINEERING AND TECHNOLOGY

GURU GHASIDAS VISHWAVIDYALAYA, BILASPUR (C.G.)

(A CENTRAL UNIVERSITY)

SESSION: 2023-24

DEPARTMENT OF ELECTRONICS AND COMMUNICATION ENGINEERING

SCHOOL OF STUDIES IN ENGINEERING AND TECHNOLOGY

GURU GHASIDAS VISHWAVIDYALAYA, BILASPUR (C.G.)

(A Central University)

CERTIFICATE

I hereby certified that the work which is being presented in the B Tech Major Project report entitled "**Deep Learning Based Face Extraction and its Enhancement For Video Surveillance** " in partial fulfillment of the requirements for the award of Bachelor of Technology in Electronics and communication Engineering and submitted to the Department of Electronics and communication Engineering, School of studies in Engineering and Technology, Guru Ghasidas Vishwavidyalaya, Central University Bilaspur, Chhattisgarh INDIA is an authentic record of my own work carried out during a period from December 2023 to May 2024 (VIII semester) under the supervision of **Jitendra Bharadwaj (Assistant professor)** ECE department .

Signature of Supervisor
Jitendra Bharadwaj
(Assistant Professor and Guide)

Head**: Dr. Soma Das**
Electronics and Communication Engineering Department

APPROVAL SHEET

This major project entitled "**Deep Learning Based Face Extraction and its Enhancement For Video Surveillance**" submitted by **Tejas Bibekar**, **Tanmay Giram** and **Nishant Wankhade** is hereby approved for the degree of Bachelor and technology in theDepartment of Electronics and communication Engineering, School of studies in Engineering and Technology, Guru Ghasidas Vishwavidyalaya, Central University Bilaspur, Chhattisgarh.

Examiners

Head of Department

Department of Electronics and communication Engineering
School of studies in Engineering and Technology,
Guru GhasidasVishwavidyalaya,
Central University, Bilaspur, Chhattisgarh.

ACKNOWLEDGEMENT

Working on this project, although was a challenge for us, would not have been achieved without the constant support, inspiration, encouragement and contribution of many people.

We are highly indebted to **Jitendra Bharadwaj (Assistant professor)** for his guidance and constant supervision as well as for providing necessary information regarding the project and also for his support in completing the project.

We owe our special thanks to our Head of Department "**Prof. Dr. Soma Das**" and Dean **Prof. Sharad Chandra Srivastava ,** School of Studies in Engineering and Technology Guru Ghasidas Vishwavidyalaya, Central University, Bilaspur (C.G.) for encouraging us to acquire courage and knowledge through this project.

We would like to express our special gratitude and thanks to our colleague in developing the project and people who have willingly helped us.

Date.....

Tejas Bibekar (20106010)
Tanmay Giram (20106021)
Nishant Wankhade (20106070)

DECLARATION

I hereby declare that the project work entitled "**Deep Learning Based Face Extraction and its Enhancement For Video Surveillance**" submitted to Guru Ghasidas Vishwavidyalaya , Central university, Bilaspur, C.G. is a record of an original work done by us under the guidance of **Jitendra Bharadwaj (Assistant professor)** , Department of Electronics and Communication Engineering, Guru Ghasidas Vishwavidyalaya , Central University, Bilaspur , Chhattisgarh.

This project work is submitted in partial fulfillment of the requirement for the award of the degree of Bachelor of Technology in Electronics and Communication Engineering.

The result embodied in this thesis have not been submitted to any other university or institute for the award of any degree or diploma.

Date .....

Sign.....
Name - Tejas Bibekar
RollNo - 20106010
Enrollment No - GGV/20/01210

Sign.....
Name - Tanmay Giram RollNo - 20106021
Enrollment No - GGV/20/01222

Sign.....
Name - Nishant Wankhade
RollNo - 20106070
Enrollment No - GGV/20/01274

Contents

List of Figures

Abstract

Surveillance systems play a crucial role in enhancing public safety and security measures. However, the effectiveness of these systems is often hindered by the poor quality of video footage, particularly when it comes to recognizing individuals of interest. Low-resolution cameras, suboptimal lighting conditions, and other environmental factors can result in captured video data that lacks the necessary clarity and detail required for accurate face recognition. This limitation poses significant challenges in investigations or situations where identifying individuals from surveillance footage is essential. To address this issue, this project aims to develop an effective pipeline for face detection, extraction, and enhancement to improve the visual quality of faces in video data. The YOLOv8 object detection model is employed to localize human keypoints within the video frames, enabling the extraction of face bounding boxes based on the detected facial keypoints. To improve the visual quality and resolution of the extracted face images, several upsampling and super-resolution techniques are explored. Initially, conventional upsampling methods are utilized. Furthermore, an advanced approach leveraging the Image Super Resolution using Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) is integrated to potentially improve the visual quality and resolution of the extracted face images. Experimental results indicate that while the ESRGAN model achieves impressive upscaling and enhancement for face images of sufficient resolution, its performance is limited when applied to low-resolution face bounding boxes extracted from videos with poor visual quality. The proposed pipeline demonstrates the potential for automated face extraction and enhancement from video data, highlighting the challenges and considerations involved in real-world applications.

Index words - Face Detection, Computer Vision, Human Keypoint Detection (YOLOv8 Pose Detection Model), Image Super-Resolution using ESRGAN implementation

1 Introduction

Video surveillance systems have become increasingly prevalent in modern society, playing a crucial role in enhancing public safety and security. However, the effectiveness of these systems is often hindered by low-resolution footage or poor visual quality, making it challenging to extract and identify individuals of interest accurately. This limitation was highlighted in a recent incident involving a theft at our residence, where the closed-circuit television (CCTV) footage captured the perpetrator's actions but failed to provide a clear and recognizable depiction of their face.

In response to this problem, our project aimed to develop a comprehensive pipeline for extracting and enhancing face images from video footage. The primary objective was to leverage state-of-the-art computer vision and deep learning techniques to improve the visual quality and resolution of face bounding boxes, thereby facilitating better identification and recognition.

Specifically, we employed the YOLOv8 pose detection model to localize human keypoints within video frames, enabling the extraction of face bounding boxes based on detected facial landmarks. Subsequently, we explored various upsampling and super-resolution techniques to enhance the resolution and visual clarity of the extracted face images. This involved experimenting with conventional methods such as Upsampling, Conv2D and Conv2DTranspose layers, as well as leveraging the advanced Image Super-Resolution using Enhanced Super-Resolution Generative Adversarial Networks (ESRGAN) pre-trained model.

While the scope of this project was limited to testing on video footage with relatively higher visual quality than the CCTV recording from the theft incident, our work aimed to demonstrate the potential of deep learning-based techniques for improving face recognition capabilities in surveillance applications. The proposed pipeline contributes to the field of computer vision by offering an automated solution for face extraction and enhancement, potentially addressing the challenges posed by low-resolution or poor-quality video data.

2 Literature Review

Face detection and recognition from video footage have been extensively studied in the field of computer vision and surveillance systems. One of the critical components of this task is the accurate localization and extraction of face regions from individual video frames. Traditionally, this has been achieved through techniques such as Viola-Jones object detection and Histogram of Oriented Gradients (HOG) descriptors. However, recent advancements in deep learning have led to the development of more robust and efficient models for object detection and pose estimation.

Pose Estimation and Keypoint Detection using YOLOv8

You Only Look Once (YOLO) is a state-of-the-art object detection algorithm that has been widely adopted for various computer vision tasks, including pose estimation. The latest version, YOLOv8, introduced by Ultralytics, incorporates advanced techniques for multi-person pose estimation, enabling the localization of up to 17 keypoints on the human body. This capability is particularly relevant for our project, as it allows for the precise extraction of face bounding boxes based on detected facial landmarks.

Pose estimation is a task that involves identifying the location of specific points in an image, usually referred to as keypoints. The keypoints can represent various parts of the object such as joints, landmarks, or other distinctive features. The locations of the keypoints are usually represented as a set of 2D `[x, y]` or 3D `[x, y, visible]` coordinates.

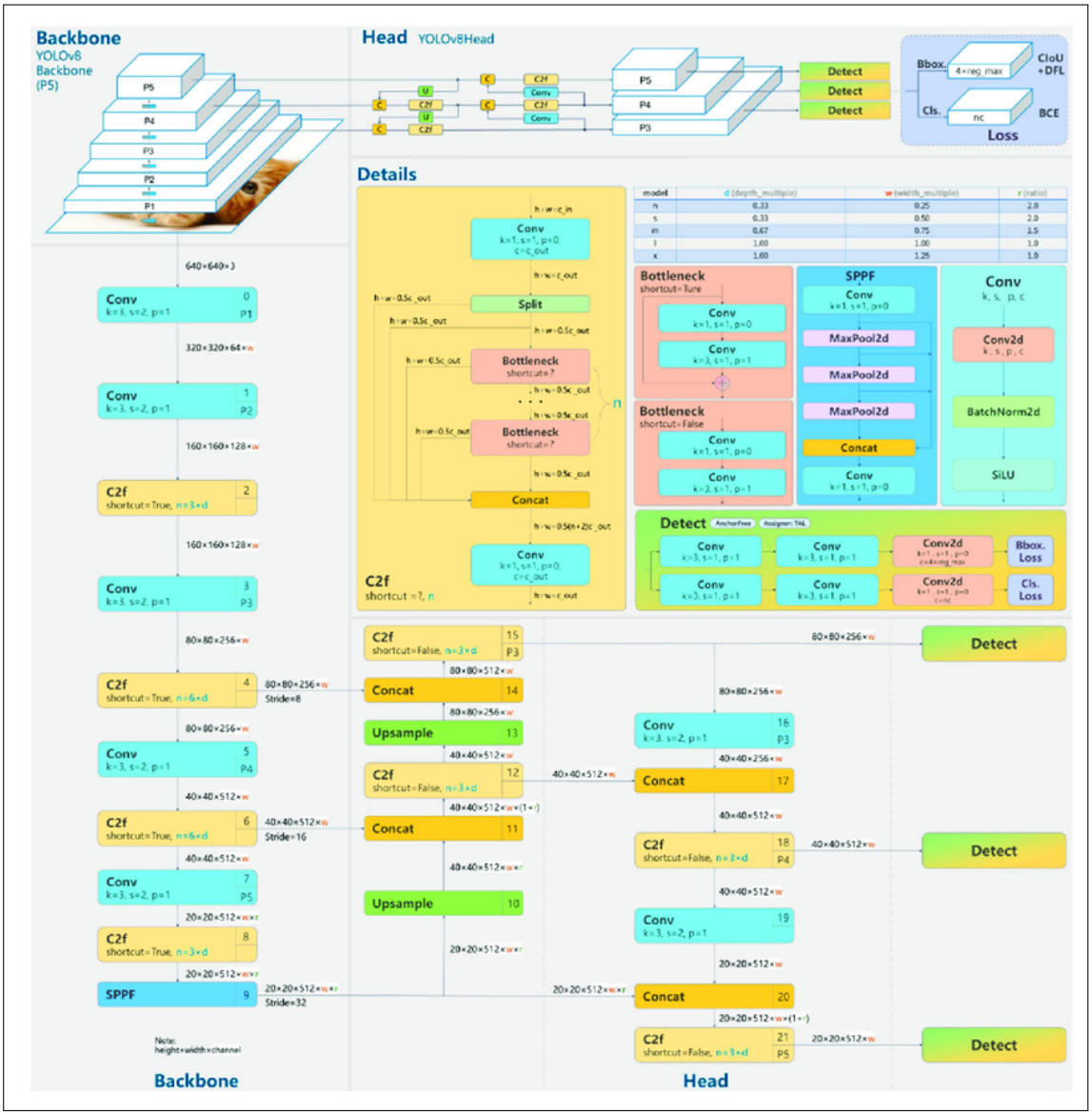


Figure 1: YOLOv8 Architecture

The YOLOv8 pose estimation model is trained on the COCO-Pose dataset, a large-scale object detection, segmentation, and pose estimation dataset derived from the popular COCO dataset. COCO-Pose focuses specifically on human pose estimation and includes multiple keypoints for each human instance. It consists of over 200,000 labeled images, providing a rich and diverse dataset for training robust pose estimation models.

Model	size (pixels)	mAP^{pose} 50-95	mAP^{pose} 50	Speed CPU ONNX (ms)	Speed A100 TensorRT (ms)	params (M)	FLOPs (B)
YOLOv8n-pose	640	50.4	80.1	131.8	1.18	3.3	9.2

Figure 2: YOLOv8 Pre-trained Nano Pose Detection Model Performance

The output of a pose estimation model is a set of points that represent the keypoints on an object in the image, usually along with the confidence scores for each point. Pose estimation is a good choice when you need to identify specific parts of an object in a scene, and their location in relation to each other.

In addition to COCO-Pose, Ultralytics provides several other datasets for pose estimation tasks, including COCO8-Pose and Tiger-Pose:

- **COCO8-Pose**: A small but versatile pose detection dataset composed of the first 8 images from the COCO train 2017 set, with 4 images for training and 4 for validation. It follows the same label format as COCO-Pose, with 17 keypoints for human poses, and is suitable for testing, debugging, and experimenting with new detection approaches.
- **Tiger-Pose**: An animal pose dataset comprising 263 images sourced from a YouTube video, with 210 images for training and 53 for validation. It follows the Ultralytics YOLO format with 12 keypoints for animal pose and no visible dimension, making it suitable for non-human pose estimation tasks.

The COCO-Pose dataset defines 17 keypoints for each human, including the nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles. The model outputs these keypoints in the Ultralytics YOLO format, representing each keypoint as a tuple of (x, y) coordinates and a confidence score for the detection. As a result, our project leverages the YOLOv8 model to accurately localize facial keypoints, which serve as the basis for extracting precise face bounding boxes from video frames.

Image Super-Resolution

While face detection and extraction are crucial steps, the enhancement of low-resolution or poor-quality face images is equally important for effective recognition and identification. Traditional image upsampling techniques, such as bicubic interpolation, often result in blurred or artifact-prone outputs. To address this limitation, deep learning-based super-resolution methods have emerged, leveraging the power of convolutional neural networks (CNNs) and Generative Adversarial Networks (GANs).

The Enhanced Super-Resolution Generative Adversarial Network (ESRGAN), proposed by Wang et al., is a state-of-the-art model for single image super-resolution. It builds upon the seminal work of the Super-Resolution Generative Adversarial Network (SRGAN) and introduces several improvements, including a Residual-in-Residual Dense Block (RRDB) architecture without batch normalization, a relativistic discriminator, and an enhanced perceptual loss function. Its network architecture is given in the Fig 3.(3). These advancements enable ESRGAN to generate realistic and natural textures while minimizing artifacts, making it a promising candidate for enhancing the visual quality of face images extracted from video footage.

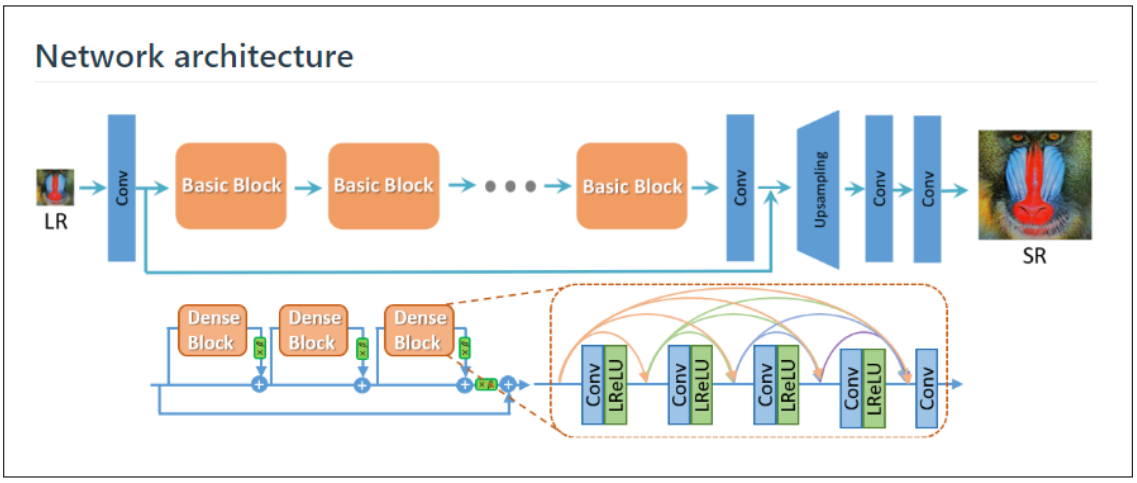


Figure 3: Network Architecture of ESRGAN

ESRGAN has demonstrated superior performance on various benchmarks, such as

Set14, BSD-500, and Urban100, outperforming previous state-of-the-art methods. Additionally, it achieved the first place in the PIRM2018-SR Challenge, further validating its effectiveness in the image super-resolution task.

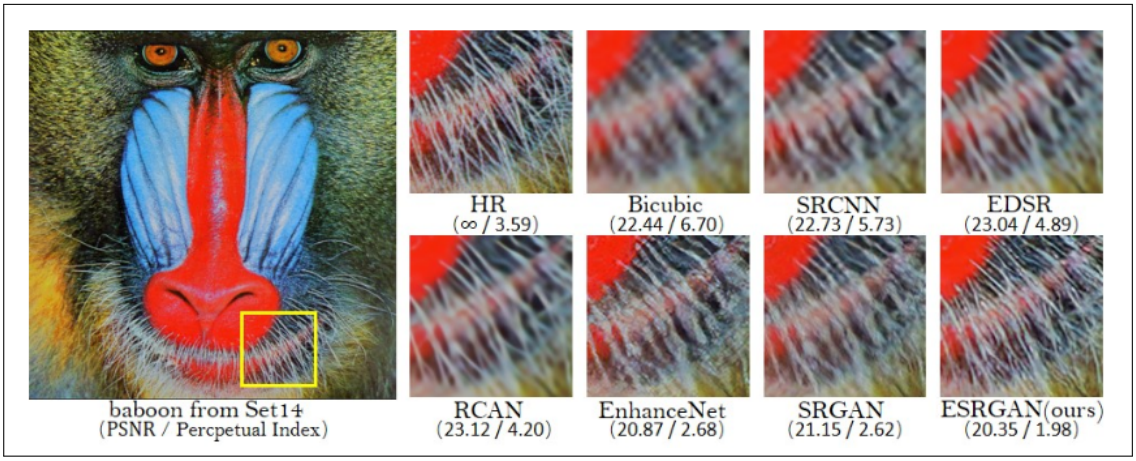


Figure 4: Quantitative Results - ESRGAN

While existing literature has explored face detection, pose estimation, and image super-resolution individually, our project aims to integrate these components into a comprehensive pipeline tailored for enhancing face images extracted from video surveillance footage. By leveraging the strengths of YOLOv8 for accurate keypoint detection and ESRGAN for high-quality super-resolution, we endeavor to address the challenges posed by low-resolution or poor-quality video data, ultimately improving the effectiveness of surveillance systems for identification and recognition purposes.

3 Methodology

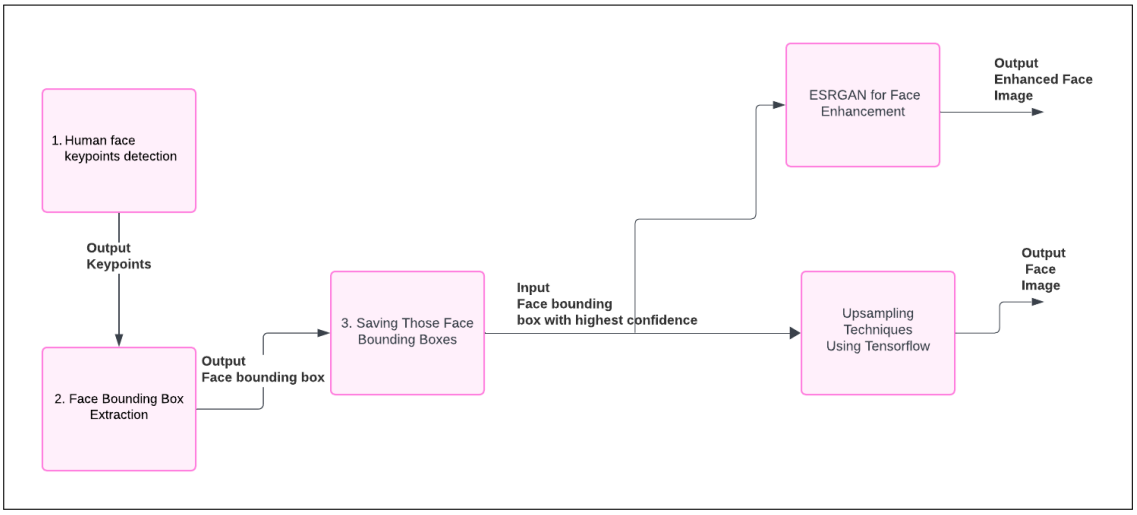


Figure 5: Flowchart

3.1 Face Extraction

The face extraction process involves several steps, starting with keypoint detection using the YOLOv8 pose estimation model, followed by post-processing techniques to localize and extract the face bounding boxes accurately.

3.1.1 Human Face keypoints Detection

1. The model takes an input of a video frame-wise and outputs a list of detection, each containing the estimated keypoints for a detected person. To handle multiple person scenarios, only the detection with the highest confidence score was considered.

2. The model provides 17 keypoints for each detection, representing various body parts such as the nose, eyes, ears, shoulders, elbows, wrists, hips, knees, and ankles. The keypoints are represented as (x, y, confidence) tuples, where x and y denote the coordinates measured from the top-left corner of the image, and confidence indicates the detection confidence for that specific keypoint.

3.1.2 Face Bounding Box Extraction and Saving them

1. From the detected keypoints, the first five keypoints (nose, left eye, right eye, left ear, right ear) were considered for face bounding box extraction.
2. The distance between the left eye and nose keypoints was calculated. If the left eye keypoint was not visible (i.e., coordinates were zero), the right eye keypoint was used instead.
3. Based on a golden ratio of 1.5, i.e. height:width ratio, derived from anthropometric studies on the relationship between facial features, the width of the face bounding box was estimated as 2 times the distance between the eye and nose keypoints. Then calculating the height on the basis of golden ratio.
4. Using the calculated width and the eye-nose distance, the top-left and bottom-right coordinates of the face bounding box were determined by subtracting and adding the width, respectively, from the eye and nose keypoint coordinates.
5. The extracted face bounding boxes were cropped from the original video frames and saved as individual image files for further processing.

This face extraction methodology leverages the accurate keypoint detection capabilities of the YOLOv8 model and applies geometric calculations based on anthropometric ratios to localize and extract face regions from video frames efficiently.

3.2 Upsampling techniques

Based on the information provided, here's an elaborated explanation of the upsampling techniques used in your methodology:

Upsampling Techniques

To enhance the resolution and visual quality of the extracted face bounding boxes, two different upsampling approaches were explored using TensorFlow's convolutional neural network (CNN) layers.

Using TensorFlow Conv2DTranspose

1. The Conv2DTranspose layer in TensorFlow is designed for upsampling tensor inputs through a fractionally-strided convolution operation.

2. A single Conv2DTranspose layer was used to upsample the input face bounding box images directly, without any additional preprocessing or resizing steps.

3. The hyperparameters for the Conv2DTranspose layer were set as follows:
   - filters=1: Single output channel, as the input and output are grayscale images.
   - kernel-size=(8, 8)‘: The size of the convolutional kernel used for upsampling.
   - ‘strides=(8, 8)‘: The stride values determine the upsampling factor, effectively increasing the spatial dimensions of the input by a factor of 8 in both height and width.
   - ‘padding="valid"‘: No padding was applied to the input tensor.
   - ‘activation=None‘: No activation function was used, as the output represents the upsampled pixel values directly.

The Conv2DTranspose layer provided a straightforward approach to upsampling the face bounding box images, but the results were not entirely satisfactory, as the upsampled images lacked fine details and exhibited blurring or artifacts.

Using TensorFlow Upsampling and Conv2D

1. To improve the upsampling quality, a combination of TensorFlow's Upsampling layer and Conv2D layers was employed.

2. The Upsampling layer was first used to increase the spatial dimensions of the input tensor.

3. Following the upsampling operation, one or more Conv2D layers were applied to the upsampled tensor

The combination of Upsampling and Conv2D layers yielded superior results compared to the Conv2DTranspose approach alone. The upsampled face bounding box images exhibited improved sharpness, reduced artifacts, and better preservation

of facial features and details.

3.3 ESRGAN for face enhancement

Both upsampling techniques were limited in their ability to recover high-frequency details and textures, especially when applied to low-resolution or poor-quality face bounding boxes extracted from surveillance footage. To further enhance the visual quality, a more advanced super-resolution technique was explored using the ESRGAN model.

Using the tensorflow implementations for Image Super Resolution, we passed the image to their implementation for enhancement.

4 Results and discussion

4.1 Dataset and Evaluation Setup

The proposed face extraction and enhancement pipeline was evaluated on a real-life scenario involving a laptop theft incident captured by a closed-circuit television (CCTV) camera. The video footage obtained from the CCTV system served as the primary dataset for testing and evaluating the effectiveness of the developed approach.

1. **Dataset Characteristics**:
    - The CCTV footage was a low-quality video, typical of routine surveillance camera systems with limited specifications.
    - No additional dataset preparation or splitting was required, as the goal was to assess the pipeline's performance on a real-world use case.
    - The video captured the incident under practical lighting conditions and environment, presenting challenges such as low resolution, motion blur, and potential occlusions.
2. **Model Setup**:
    - The YOLOv8 pose estimation model and the ESRGAN super-resolution model were employed in their pre-trained state, leveraging their robust capabilities and state-of-the-art performance.
    - No additional training or fine-tuning was performed on these models, as they were designed to generalize well to diverse scenarios.

4.2 Face Extraction Performance

The YOLOv8 pose estimation model demonstrated remarkable accuracy in facial keypoint detection, even when using the lightweight nano variant. Despite the low-resolution nature of the CCTV footage, the model successfully localized facial keypoints, enabling the precise extraction of face bounding boxes from the video frames.

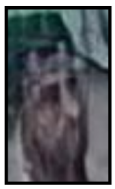
Input Frame

Output after Pose Detection

Input Face

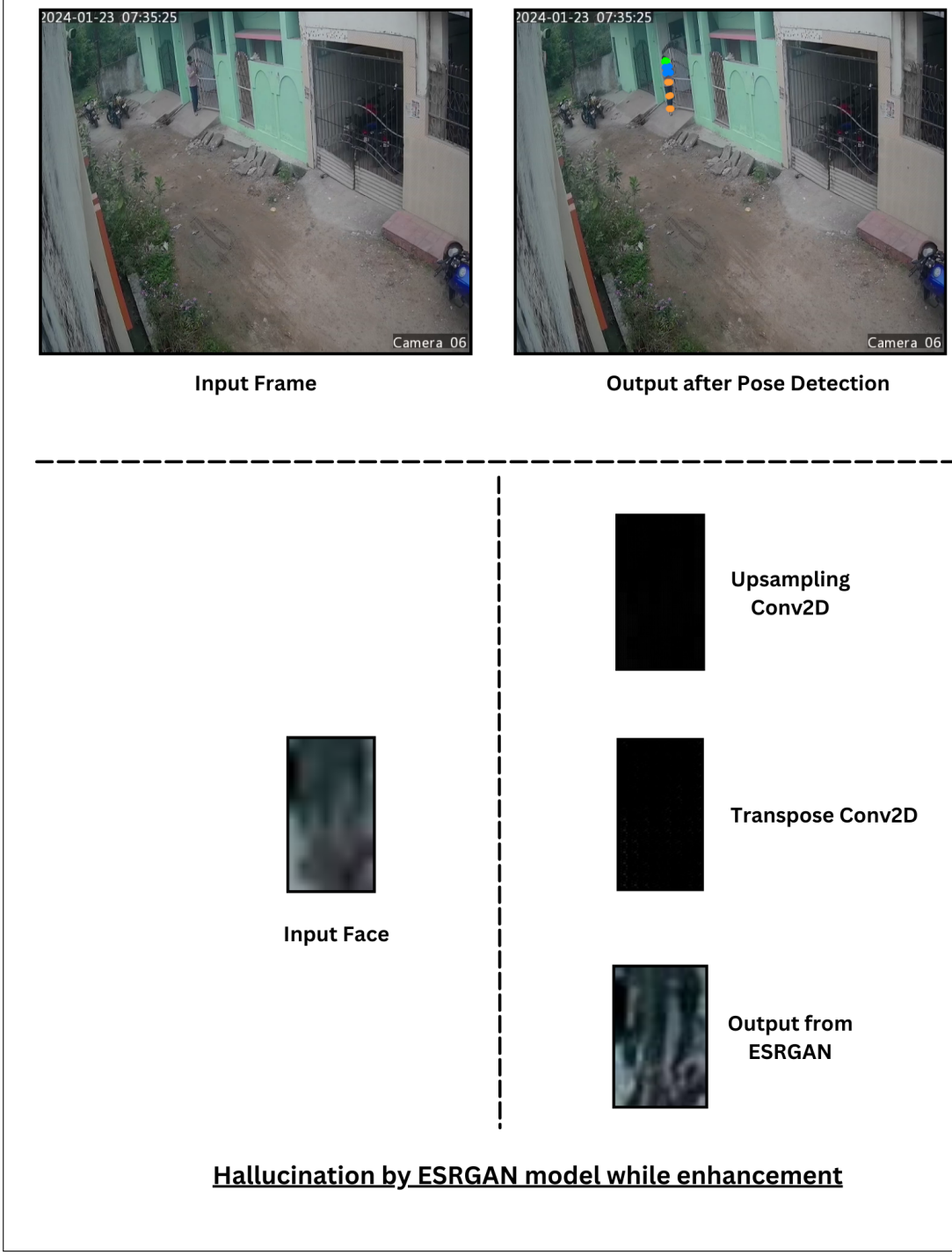
Upsampling
Conv2D

Transpose Conv2D



Output from
ESRGAN

Better Enhancement by ESRGAN


2024-01-23 07:35:25
Camera 06
Input Frame
2024-01-23 07:35:25
Camera 06
Output after Pose Detection
Upsampling Conv2D
Transpose Conv2D
Output from ESRGAN
Input Face
Hallucination by ESRGAN model while enhancement

The success rate of face bounding box extraction was satisfactory, allowing the pipeline to reliably capture and extract face images from most of the video frames. Even in challenging conditions with low-resolution inputs, the model's performance remained robust, highlighting its effectiveness in real-world surveillance scenarios.

4.3 Discussion

The proposed pipeline demonstrated promising results in extracting and enhancing face bounding boxes from challenging CCTV footage. The combination of YOLOv8 for keypoint detection and subsequent bounding box extraction allowed for accurate face localization, even in low-resolution environments. However, the enhancement techniques, including conventional upsampling methods and the ESRGAN model, exhibited limitations when dealing with extremely small or degraded face images.

A key limitation of the current pipeline is its reliance on reasonably good quality video feeds. While it performed well on the CCTV footage, it may face challenges with severely degraded or low-resolution videos, where face bounding boxes lack sufficient detail for effective enhancement.

Despite these limitations, the pipeline holds practical implications in surveillance systems, forensic investigations, and law enforcement. By enhancing the visual quality and resolution of face images, it can potentially improve the accuracy and reliability of subsequent face recognition tasks, aiding in the identification of individuals of interest.

The modular nature of the pipeline allows for the integration of alternative or improved techniques as new advancements emerge. Future work could explore advanced face extraction methods, specialized super-resolution models for surveillance applications, or additional preprocessing/post-processing steps to address real-world challenges.

Overall, the proposed pipeline leverages state-of-the-art deep learning models and computer vision techniques to tackle the problem of face extraction and enhancement from low-quality video sources. While exhibiting limitations, it serves as a

foundation for further research and development, contributing to improved public safety and security measures through enhanced face recognition capabilities.

5 Conclusions

The project successfully developed a pipeline for face detection, extraction, and enhancement using computer vision techniques. The YOLOv8 object detection model was employed to localize human keypoints within the video frames, enabling the extraction of face bounding boxes based on the detected facial keypoints. The project explored various upsampling and super-resolution techniques, including conventional methods and an advanced approach using the ESRGAN pre-trained model. The experimental results indicated that the ESRGAN model achieved impressive upscaling and enhancement for face images of sufficient resolution but was limited when applied to low-resolution face bounding boxes extracted from videos with poor visual quality.

However, the enhancement techniques employed, while capable of improving the visual quality and resolution of the extracted face bounding boxes, exhibited limitations when dealing with extremely small or degraded face images. The conventional upsampling methods and the ESRGAN model performed optimally when applied to face bounding boxes with sufficient resolution and quality, but their effectiveness diminished as the input quality deteriorated. A key limitation of the current pipeline is its reliance on the availability of a reasonably good quality video feed. While the approach demonstrated promising results on the CCTV footage, it may encounter challenges when dealing with severely degraded or low-resolution videos, where the face bounding boxes become too small or lack sufficient detail for effective enhancement.

However, it is worth noting that the proposed pipeline may still work to some extent with poor quality videos, as the ESRGAN model can still improve the visual quality of face images to a certain extent, even if the results are not as impressive as with higher quality videos. Future work could explore more advanced face extraction methods, specialized super-resolution models tailored for surveillance applications, or the incorporation of additional preprocessing or post-processing steps to address specific challenges encountered in real-world scenarios.

6 Future Scope

Emotion Recognition:

Utilizing enhanced facial images for more accurate emotion recognition systems has diverse applications across various fields. In market research, for instance, more precise emotion recognition can provide valuable insights into consumer preferences and behavior. In mental health, accurate emotion recognition can assist therapists and clinicians in assessing patients' emotional states and tracking changes over time. Moreover, in human-computer interaction, enhanced emotion recognition capabilities can enable more natural and intuitive interactions with devices and interfaces.

Biometric Authentication:

Enhancing the quality of facial images used in biometric authentication systems can improve the accuracy and reliability of identification and authentication processes. Enhanced facial images can lead to more robust biometric recognition systems, reducing the risk of false positives and false negatives. This application has implications for various sectors, including access control, border security, and financial transactions, where secure and reliable authentication methods are essential.

7 References

References

[1] Ultralytics. (n.d.). Pose Estimation. Retrieved from <https://docs.ultralytics.com/tasks/pose/>

[2] Ultralytics. (n.d.). Pose Datasets Overview. Retrieved from <https://docs.ultralytics.com/datasets/pose/>

[3] TensorFlow Hub. (n.d.). Image Super Resolution using ESRGAN TensorFlow Hub. Retrieved from https://www.tensorflow.org/hub/tutorials/image enhancing

[4] ESRGAN. (n.d.). Network Architecture. Retrieved from https://esrgan.readthedocs.io/en/latest/pages/esrgan.html - network-architecture

[5] TensorFlow Hub. (n.d.). ESRGAN. TensorFlow Hub. Retrieved from https://tfhub.dev/tensorflow/esrgan x4/1

[6] Zhang, Y., Liu, X., and Wang, Z., 2020, Image Super-Resolution using Enhanced Super-Resolution Generative Adversarial Networks, arXiv preprint arXiv:2004.05983.