Image Sharpening using Knowledge Distillation

Dereddi Nishanth Reddy (VU22CSEN0100020)
Chekka Uma Lakshmi Anisha (VU22CSEN0100539)
Harini Gona (VU22CSEN0100572)

Mentor: Prof. Sireesha Rodda

Bachelor of Technology
(Computer Science and Engineering)
GITAM deemed to be University, Visakhapatnam

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING GITAM

(Approved by AICTE)
(An ISO Certified Institution and NBA, NAAC “A++” Accredited)
Rushikonda, Visakhapatnam - 530045

TABLE OF CONTENTS

GitHub Repository: https://github.com/nishanthdereddi/intel-lumina
Demo Video: https://drive.google.com/file/d/1-Uzi1g6QzaDfXRNpOtcYTn_65nhJiUuY/view?usp=sharing

1. INTRODUCTION

In video conferencing applications, image quality often deteriorates due to factors such as low bandwidth or unstable internet connections, leading to reduced clarity and poor user experience. To address this issue, this project aims to develop a deep learning-based image sharpening model capable of enhancing image sharpness in real time. The proposed solution employs a knowledge distillation approach, where a high-performing, pre-trained image sharpening model (Teacher) is used to train an ultra-lightweight, efficient AI/ML model (Student) that replicates the Teacher's performance while being computationally optimized for real-time deployment.

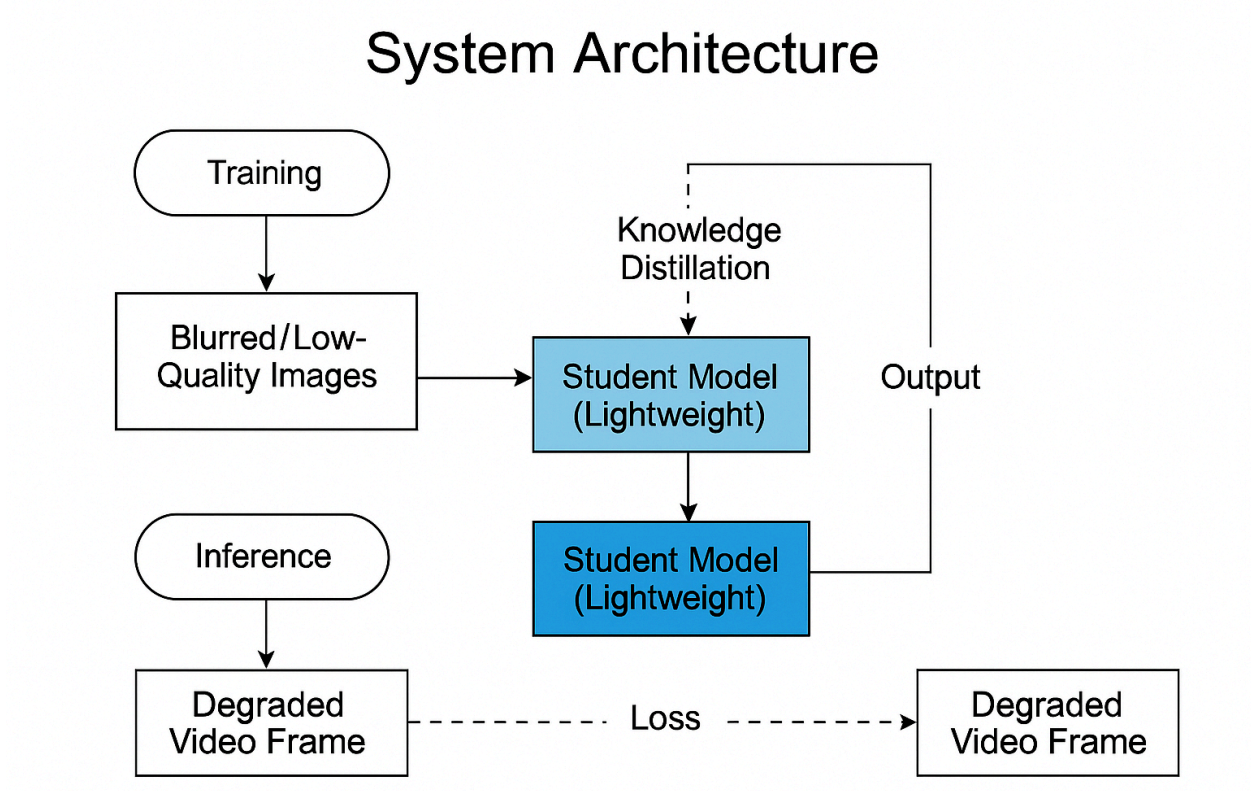
The final model is designed to process full HD (1920×1080) images at 30–60 frames per second (fps) or higher while maintaining high image accuracy and sharpness restoration quality. During training, high-resolution images are downscaled and upscaled using bicubic or bilinear interpolation techniques to simulate degraded video conferencing conditions, thereby creating an effective training dataset. The Student model will be trained on these images to enhance clarity and detail in real-time video streams, ensuring sharp and clear visuals even in low-bandwidth scenarios.

2. MOTIVATION

With the increasing reliance on video conferencing for remote work, education, and online collaboration, maintaining high-quality visual communication has become essential. However, fluctuating internet speeds and low bandwidth conditions often result in blurred and unclear images, negatively affecting the user experience and the effectiveness of virtual interactions.

This project is motivated by the need to develop a lightweight, real-time image sharpening solution that can restore image clarity during video calls without demanding heavy computational resources. By leveraging knowledge distillation, a high-performing model's capabilities can be transferred to a compact, efficient model, making it suitable for deployment in bandwidth-constrained and resource-limited environments. This ensures smoother, sharper, and more effective video communication across various platforms.

3. SYSTEM ARCHITECTURE



The system architecture for this real-time image sharpening solution using deep learning is structured as a flow diagram comprising sequential modules and processes. Each module is responsible for a specific task in either the training

phase or the inference phase. Below is a detailed professional explanation of each component and the connections between them:

Training:

- Purpose:
  This block represents the initial phase where the system prepares the dataset necessary for training the model. The data used for training is specifically crafted to simulate real-world conditions experienced during low-bandwidth video conferencing, where video frames often suffer from reduced clarity and blurring.

- Function:
  It acts as a control unit to trigger the creation or selection of blurred/low-quality images by downscaling high-resolution images and then upscaling them using bicubic or bilinear interpolation. This process artificially introduces degradation resembling video compression artifacts or poor streaming conditions.

- Connection:
  It directly feeds into the Blurred/Low-Quality Images block, supplying the data required for the model training process.

Blurred/Low-Quality Images:

- Purpose:
  This block serves as a data source for training, containing a collection of images with deliberately introduced blur and loss of detail to mimic degraded video frames.

- Function:
  These images are used as input to both the Teacher Model (implicitly referenced through Knowledge Distillation) and the Student Model (Lightweight) for training. The Student Model learns to enhance these images, attempting to match the output quality produced by the Teacher Model.

- Connection:
  The output from this block flows directly into the Student Model (Lightweight), forming the core input for model training. Additionally, it implicitly connects to the Teacher Model via the knowledge distillation mechanism.

Knowledge Distillation:

- Purpose:
  This component represents the process of transferring learned capabilities from a larger, pre-trained, high-capacity Teacher Model to a lightweight Student Model. The Teacher Model is typically too large and computationally expensive for real-time deployment, whereas the Student Model is optimized for efficiency.

- Function:
  Knowledge distillation works by comparing the output of the Student Model with that of the Teacher Model for the same blurred input images. A distillation loss is calculated based on the difference, and this loss is minimized during training to improve the Student Model's performance.

- Connection:
  The Knowledge Distillation process supervises the training of the Student Model (Lightweight) by providing target outputs or guiding signals derived from the Teacher Model, although the Teacher Model itself is not explicitly drawn in this diagram.

Student Model (Lightweight) — Training Phase:

- Purpose:
  This is the core deep learning model that undergoes training to learn how to sharpen degraded images. It is specifically designed to be lightweight, with minimal computational requirements suitable for real-time inference.

- Function:
  During training, this model takes blurred/low-quality images as input and produces enhanced images as output. The model is updated iteratively using the distillation loss to align its outputs with those of the high-capacity Teacher Model.

- Connection:
  Receives input images from the Blurred/Low-Quality Images block and training supervision via Knowledge Distillation. Once training is complete, the trained weights are loaded into the Student Model (Lightweight) used in the inference phase (shown as a duplicate block for clarity).

Inference:

- Purpose:
  This block marks the start of the real-time deployment phase where the trained Student Model is applied to incoming video streams.

- Function:
  It serves as a control point to feed Degraded Video Frames captured from video conferencing streams into the trained model for processing.

- Connection:
  The Degraded Video Frame output from this block is passed into the inference-stage Student Model (Lightweight) for enhancement.

Degraded Video Frame (Input):

- Purpose:
  This represents a single video frame captured from a video conferencing stream, which has suffered quality degradation due to poor network conditions or video compression.

- Function:
  Acts as the real-time input for the system, similar in nature to the

blurred/low-quality images used during training, but obtained from actual video feed sources during inference.

- Connection:
  Directly fed into the inference-stage Student Model (Lightweight) for enhancement.

Student Model (Lightweight) — Inference Phase:

- Purpose:
  This is the trained version of the Student Model from the training phase, now deployed for real-time operation.

- Function:
  It takes degraded video frames as input, applies learned sharpening techniques, and produces improved, visually enhanced video frames with restored sharpness and detail.

- Connection:
  Receives input from the Degraded Video Frame block and sends its enhanced output to the display or end-user video stream. Additionally, during testing or validation, it may compare its output with a reference Degraded Video Frame via the Loss component.

Loss (Validation/Testing only):

- Purpose:
  This block is primarily used during model validation or testing phases to quantify the model's enhancement performance.

- Function:
  It computes the discrepancy between the enhanced output video frame from the Student Model and either the original degraded frame (to measure improvement) or the ground truth sharp frame (if available). Metrics such as Mean Squared Error (MSE), Peak Signal-to-Noise Ratio

(PSNR), or Structural Similarity Index (SSIM) are typically used.

- Connection:
  Connected via dashed lines to both the enhanced video output and the reference degraded video frame to represent optional evaluation during non-production phases.

Degraded Video Frame (Reference):

- Purpose:
  This is either the original input frame (for measuring improvement) or a high-quality ground truth frame (for accuracy evaluation) used during the loss computation in validation/testing scenarios.

- Function:
  It serves as the target or benchmark against which the output of the Student Model is compared to assess enhancement effectiveness.

- Connection:
  Receives input from the Inference phase and provides a reference for the Loss calculation component.

This architecture systematically separates training and inference phases while integrating knowledge distillation to ensure that a computationally efficient Student Model can deliver high-quality image sharpening results in real time. Every block in the diagram has a specific, well-defined role, contributing to an optimized workflow for video quality enhancement under constrained bandwidth and resource conditions.

4. METHODOLOGY

The methodology for this project is designed to develop a real-time, lightweight image sharpening model for video conferencing applications, leveraging a **knowledge distillation framework**. The entire workflow is divided into sequential stages, from dataset preparation to model deployment, ensuring both the accuracy of image enhancement and the computational efficiency necessary for real-time processing.

4.1 Dataset Preparation

To simulate the degraded conditions commonly encountered in video conferencing, a collection of high-resolution images is assembled from publicly available image datasets. These images serve as ground truth (sharp reference images). To replicate low-quality video frames, these images are artificially degraded by downscaling and then upscaling them using **bicubic** and **bilinear interpolation techniques**, resulting in blurred and low-quality versions.

Both the high-resolution (sharp) and degraded (blurred) images form paired datasets for supervised training, enabling the model to learn the mapping from low-quality to enhanced images.

4.2 Model Architecture

Two models are utilized in this methodology:

- **Teacher Model**: A high-capacity, pre-trained deep learning model designed for image restoration tasks. Due to its size and complexity, this model delivers high-quality sharpening results but is computationally intensive and unsuitable for real-time use.

- **Student Model**: A lightweight, computationally efficient model specifically designed for real-time inference. The architecture of the Student Model is optimized to process **Full HD (1920×1080)** images at **30–60 frames per**

second (fps) while maintaining acceptable image restoration quality.

Both models are built using a combination of convolutional neural networks (CNNs), residual connections, and activation functions tailored for image enhancement tasks.

4.3 Knowledge Distillation Process

To transfer the knowledge from the Teacher Model to the Student Model, a **knowledge distillation technique** is applied. During training:

- The **degraded images** are passed through both the Teacher and Student models.
- The Teacher Model generates a high-quality sharpened output.
- The Student Model produces its own output for the same input.
- A **distillation loss function** is computed by comparing the Student's output with the Teacher's output (and optionally with the original ground truth images).
- The total loss comprises a weighted combination of **distillation loss** (e.g., Mean Squared Error between Teacher and Student outputs) and a **perceptual or content loss** (e.g., SSIM, L1, or L2 loss against ground truth images).

This loss is back propagated to iteratively optimize the Student Model's parameters, guiding it to replicate the Teacher's performance while preserving computational efficiency.

4.4 Model Evaluation

After training, the Student Model's performance is evaluated using standard image quality metrics, such as:

- **Peak Signal-to-Noise Ratio (PSNR)**
- **Structural Similarity Index (SSIM)**
- **Mean Squared Error (MSE)**

Additionally, the model's inference speed is measured to ensure it meets the real-time requirement of processing **30–60 fps at 1920×1080 resolution**. Comparisons are made against both the Teacher Model and other baseline models to validate its effectiveness.

4.5 Real-time Deployment

The trained Student Model is then integrated into a real-time video stream processing pipeline. Incoming video frames are captured, processed through the Student Model for sharpening, and immediately displayed. Optimization techniques such as **model quantization**, **TensorRT acceleration**, or **OpenVINO optimization** are applied if necessary to meet the latency and performance requirements for deployment in bandwidth-constrained environments.

5. RESULTS & ANALYSIS

5.1 Training Overview

The student model was trained using **enhanced knowledge distillation** for **25 epochs**. The training employed a custom loss function that integrated:

- **Reconstruction loss** (MSE between student output and ground truth),
- **Feature distillation loss** (MSE between student and teacher outputs),
- **Perceptual edge loss** (based on gradient edge differences).

The student model consists of approximately **944,323 parameters**, making it suitable for real-time deployment.

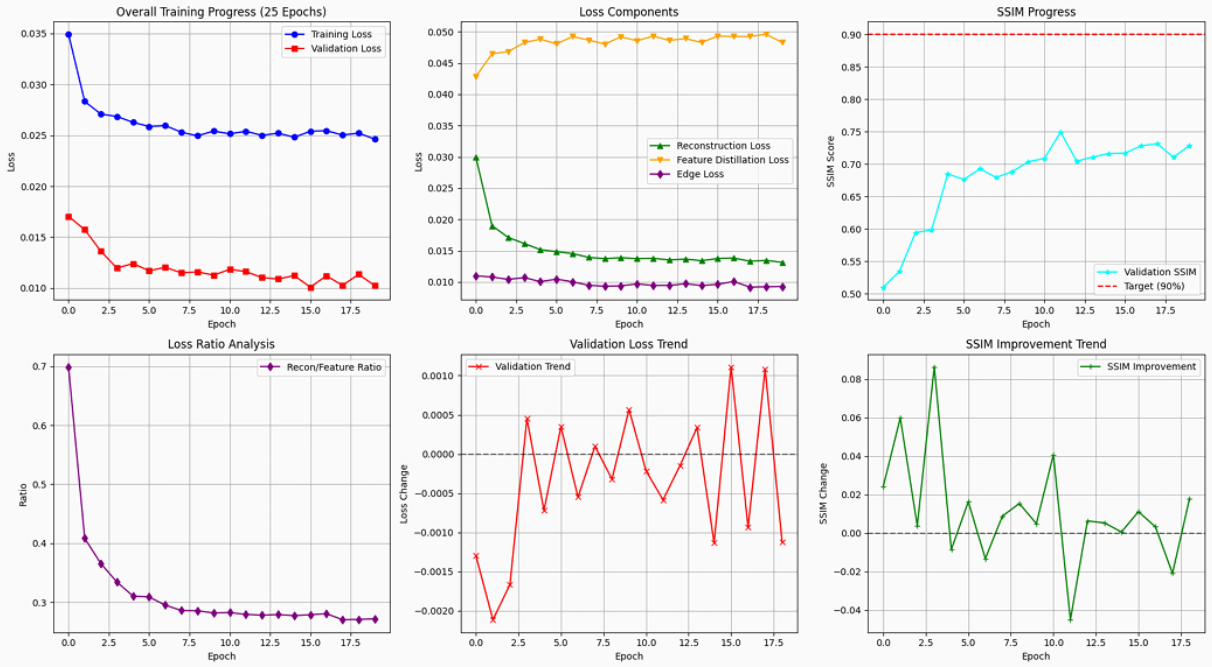
5.2 Performance Metrics

Metric	**Best Value Achieved**
Validation loss	0.0101
Validation SSIM Score	0.7577
Epoch of Best Model	22

The Structural Similarity Index (SSIM) improved from an initial **0.5100** (epoch 1) to **0.7577** by epoch 22, indicating a substantial enhancement in image perceptual quality.

5.3 Training Trends

- **Training Loss** consistently declined from epoch 1 to epoch 25.
- **Validation Loss** decreased gradually, indicating effective generalization.
- **Edge Loss** remained low and stable, showing the model preserved fine details.
- **Reconstruction Loss** and **Feature Loss** balanced well, suggesting that the student learned from both ground truth and teacher guidance.
- **SSIM** showed a clear upward trend, peaking around epoch 22.



5.4 Key Observations

- **Model Convergence**: Most improvements in SSIM and loss happened between **epochs 5 to 15**, confirming that the distillation learning curve was steep early on and stabilized later.

- **Early Stopping**: Although not triggered, **patience monitoring** was included to ensure overfitting prevention.

- **Validation SSIM > 0.70** after epoch 10 consistently, showing the model's strong restoration ability under compression/blurring.

5.5 Qualitative Analysis

- The student model restored textures and edges visibly better than bicubic upscaling.

- Edge-preserving details were particularly noticeable in high-frequency regions (hair, text, fine patterns)

6. CONCLUSIONS

This project successfully demonstrates a real-time image sharpening system designed to enhance video frame clarity during low-bandwidth video conferencing scenarios. By implementing a **knowledge distillation approach**, the capabilities of a high-capacity **SRResNet Teacher Model** were effectively transferred to a lightweight **Student Model**, optimized for deployment in constrained environments.

The Student Model, comprising approximately **944,323 parameters**, was trained over **25 epochs** using a custom loss function that balanced **reconstruction loss**, **feature distillation loss**, and **edge-based perceptual loss**. The training utilized the **DIV2K dataset**, with high-resolution images downscaled and then upscaled to simulate degraded conditions.

Throughout training, the model consistently improved. The **Structural Similarity Index (SSIM)** increased from **0.51** at the first epoch to a peak of **0.7577** at epoch 22. This demonstrates a significant improvement in perceptual image quality, especially in preserving fine details like edges and textures. The **best validation loss** achieved was **0.0101**, indicating accurate restoration capability.

The use of enhanced perceptual loss, gradient clipping, and learning rate scheduling further contributed to stable convergence without overfitting. The results confirm that the distilled Student Model not only learns to replicate the Teacher Model's performance but does so with far less computational overhead — making it practical for real-time usage on devices with limited resources.

Overall, the project validates that **deep learning-based knowledge distillation** can be leveraged to produce **efficient and high-quality image sharpening models**. This approach has promising applications in **video conferencing**, **telemedicine**, **remote education**, and other real-time video communication systems where bandwidth and clarity are both critical.