Face Super-Resolution and Occlusion-Robust Face Recognition: A Unified Deep Learning Framework

Shrayash Barnwal

Contents

1 Introduction

Face recognition has become a core component of modern security, authentication, and human–computer interaction systems. From mobile device unlocking and access control to video surveillance and forensics, the reliability of these systems depends critically on their ability to recognize faces under diverse and challenging real-world conditions.

Deep convolutional neural networks (CNNs) and margin-based embedding losses have led to major advances in face recognition performance. However, even state-of-the-art models struggle under two ubiquitous degradations:

- **Low-resolution inputs**, such as faces captured from distant cameras, legacy CCTV footage, or extreme downsampling.
- **Facial occlusions**, such as medical masks, sunglasses, scarves, hair, and handheld objects.

Low resolution removes fine-grained identity cues, while occlusions selectively hide or distort key regions such as the eyes, nose, or mouth. In real deployments, both problems often occur simultaneously, leading to substantial performance drops.

This project brings together three consecutive stages of work:

1. Early attempts to train a deep CNN for face recognition from scratch on limited data.
2. A dedicated study of face super-resolution (SR) using state-of-the-art models, evaluated with standard image quality metrics across resolutions.
3. An implementation of an occlusion-robust recognition pipeline, inspired by the Adaptive Multi-Type Occlusion-Aware Face Recognition Model (AMOFR) framework, which generates synthetic occlusions and uses a MobileFaceNet + ArcFace backbone for recognition.[1]

The goal is to understand not only how each component behaves in isolation, but also how super-resolution and occlusion-aware training interact in the broader face recognition pipeline.

1.1 Motivation

From a practical viewpoint, several motivating scenarios justify this work:

- CCTV or campus surveillance cameras that capture faces at very low resolution, where raw inputs are too small for reliable recognition.

[1] Y. Liu *et al.*, "Adaptive Face Recognition for Multi-Type Occlusions," IEEE TCSVT, 2024.

- Post-pandemic environments where masks, glasses, and other accessories have become commonplace, significantly altering facial appearance.
- Legacy datasets or archival footage where re-capturing data at higher quality is not possible, so software-based enhancement is the only option.

In such settings, simply applying a standard deep face recognizer is insufficient. We require methods that can:

1. enhance facial images in a way that preserves identity, and
2. train recognition models to be robust to partial visibility caused by occlusions.

1.2 Problem Statement

The core problem addressed in this report can be phrased as:

> *How can we improve deep face recognition performance when faces are both low-resolution and affected by multi-type occlusions, using a combination of super-resolution and occlusion-aware training?*

This study focuses on the verification setting, where the task is to decide whether two face images correspond to the same person, under varying degrees of degradation.

1.3 Objectives

The main objectives of the project are:

- To benchmark multiple face SR models at different input resolutions and analyze how reconstruction metrics behave across models and scales.
- To adapt and implement an AMOFR-style occlusion synthesis and training pipeline and evaluate its behaviour through training curves and verification metrics.
- To connect the findings from the SR and occlusion components and discuss how they should be combined in a practical recognition system.

2 Background

2.1 Deep Face Recognition and Embedding Spaces

Modern face recognition systems map an input face image to a compact embedding vector in a high-dimensional space. Two images of the same person should lie close together in this space, while embeddings of different people should be well separated.

Given an input image I, a network f_θ with parameters θ outputs a feature vector

$$x = f_\theta(I) \in \mathbb{R}^d.$$

Embeddings are usually L2-normalized, $\hat{x} = x/\|x\|_2$, so that cosine similarity reduces to a dot product:

$$\mathrm{sim}(\hat{x}_1, \hat{x}_2) = \hat{x}_1^\top \hat{x}_2.$$

Classical softmax cross-entropy loss enforces correct classification but does not explicitly control the angular separation between classes. Margin-based losses such as ArcFace introduce an angular margin between classes on the hypersphere, encouraging tighter intra-class clusters and larger inter-class gaps—a property that becomes particularly important when input images are degraded.

2.2 Earlier CNN Experiments

In the earliest phase of this project, a custom CNN architecture was trained from scratch on limited subsets of face datasets. These experiments, documented in the SOP report, showed typical symptoms of overfitting:

- training accuracy improved steadily while validation accuracy plateaued at a relatively low level;
- the model was very sensitive to minor changes in hyperparameters and data splits;
- small variations in preprocessing caused noticeable performance differences.

These results indicated that, given the available data and compute, it would be more effective to focus on improving the quality and robustness of inputs (via SR and occlusion modeling) rather than attempting to design a large standalone recognition network from scratch. They also highlighted how strongly recognition performance depends on image quality and on the distribution of training data.

2.3 Face Super-Resolution

Face super-resolution aims to reconstruct a high-resolution (HR) face image from a low-resolution (LR) input. Unlike generic image SR, face SR must preserve identity-specific details that are critical for recognition, such as:

- relative distances between facial landmarks,
- shape of eyes, nose, and mouth,

- fine-grained texture patterns like wrinkles and eyebrows.

Given an LR image I_{LR} and its HR counterpart I_{HR}, a super-resolution model G_ϕ learns a mapping

$$\hat{I}_{SR} = G_\phi(I_{LR}),$$

where $\hat{I}_{SR}$ is the reconstructed image. Performance is typically evaluated by:

- **PSNR (Peak Signal-to-Noise Ratio)**,

$$\text{PSNR} = 10 \log_{10} \left(\frac{MAX^2}{\text{MSE}(I_{HR}, \hat{I}_{SR})} \right),$$

  where MAX is the maximum possible pixel value.

- **SSIM (Structural Similarity Index)**, which compares luminance, contrast, and structure between I_{HR} and $\hat{I}_{SR}$.

- **LPIPS**, a learned perceptual similarity measure computed from deep network features; lower values indicate greater perceptual similarity.

Super-Resolution Models Considered

In this project, four SR models are compared:

- **GFPGAN** — a generative model that leverages a pretrained GAN prior to reconstruct plausible facial textures, often producing visually appealing but sometimes identity-altering results at extreme degradations.

- **Real-ESRGAN** — an enhanced SRGAN variant trained with a more realistic degradation model, designed to handle unknown blur, noise, and compression artifacts.

- **SwinIR** — a transformer-based SR model that uses shifted window attention to capture long-range dependencies and restore high-frequency details.

- **BSRGAN** — an SR method that focuses on realistic degradation simulation, often producing natural-looking textures in the presence of complex distortions.

Each model is evaluated on faces downsampled to three resolutions: 16×16, 32×32, and 64×64 pixels. The emphasis is on how well the models restore facial structure and texture at each resolution.

2.4 Face Recognition With Occlusions and AMOFR

Occlusions such as masks, glasses, and sunglasses hide parts of the face and can cause deep models to misinterpret or ignore essential features. Standard recognition models trained only on unoccluded faces often perform poorly when tested on occluded faces.

The AMOFR framework proposes an adaptive multi-type occlusion-aware face recognition model with three main components:

1. **Occluded face image generator** that simulates masks, glasses, and sunglasses using facial landmarks.
2. **Knowledge distillation** from a teacher model trained on unoccluded faces to a student model trained on occluded faces.
3. **Occlusion type-based adaptive module** that uses a CLIP-based occlusion classifier and dynamic weighted fusion to reduce intra-class variations caused by different occlusion types.

Figure 1 shows the AMOFR framework as proposed in the original paper.

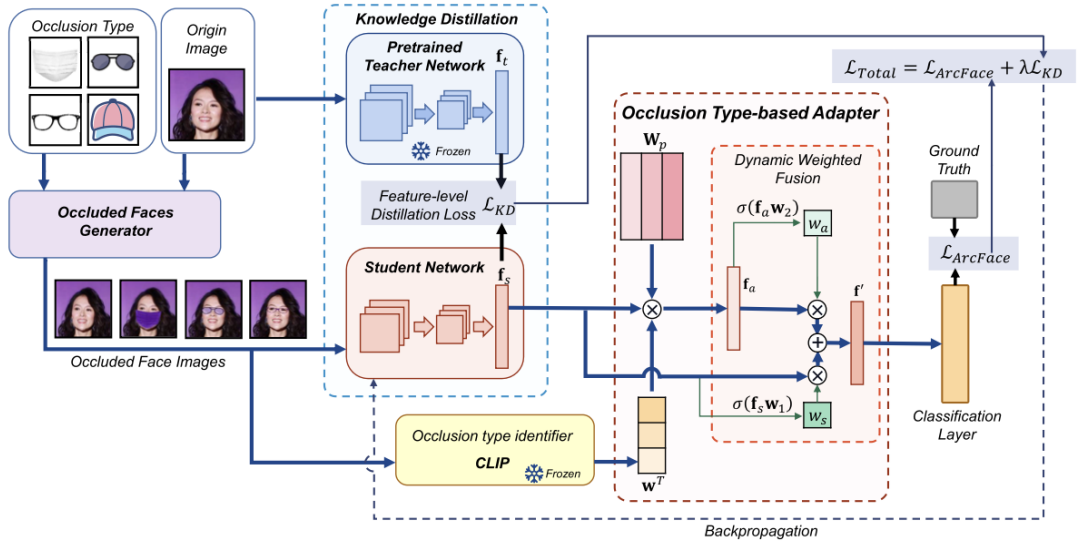

Figure 1: AMOFR framework with occlusion generator, teacher–student knowledge distillation, and occlusion-type adapter. (From the AMOFR paper.)

The key takeaway is that AMOFR does not try to "complete" or inpaint occluded regions. Instead, it learns to extract stable features from the visible parts of the face and to normalize variations introduced by different occlusion types through adaptive fusion.

2.5 Occlusion Types and Taxonomy

From the perspective of recognition, different occlusions have different impacts:

- **Rigid occlusions** such as sunglasses and shields cover a region with a solid object, blocking most texture.
- **Non-rigid occlusions** such as cloth masks and scarves deform with facial movement and can create complex boundaries.
- **Transparent occlusions** such as regular glasses partially preserve texture but introduce reflections and refractions.
- **Local occlusions** (hands, hair strands) often cover only part of the face and may appear in unusual locations.

Models trained without explicit exposure to these variations are likely to overfit to unoccluded faces and fail in real use cases. This motivates synthetic occlusion generation as a central tool in AMOFR and in this project.

3 Methodology

This section describes the overall methodology used in the project, covering datasets, preprocessing, the super-resolution workflow, occlusion synthesis pipeline, and the recognition model.

3.1 Datasets and Pair Construction

The experiments primarily use:

- **Labeled Faces in the Wild (LFW)** for both SR evaluation and recognition testing.
- Small subsets of identities for early CNN experiments and for SR comparisons, as documented in previous reports.

Faces are evaluated in a verification setting: given a pair of images (I_1, I_2), the system must decide whether they belong to the same person. Positive pairs contain two images of the same identity; negative pairs contain images from different identities.

For a given decision threshold τ on cosine similarity, the following outcomes are defined:

- **True Positive (TP):** same-identity pair with similarity $\geq \tau$.
- **False Positive (FP):** different-identity pair with similarity $\geq \tau$.
- **True Negative (TN):** different-identity pair with similarity $< \tau$.

- **False Negative (FN):** same-identity pair with similarity $< \tau$.

From these we can compute accuracy, true positive rate (TPR), and false positive rate (FPR) as a function of τ:

$$\text{Accuracy}(\tau) = \frac{TP + TN}{TP + TN + FP + FN},$$

$$\text{TPR}(\tau) = \frac{TP}{TP + FN}, \qquad \text{FPR}(\tau) = \frac{FP}{FP + TN}.$$

These definitions underlie the accuracy-vs-threshold and ROC curves shown later.

3.2 Preprocessing Pipeline

To standardize inputs before SR or occlusion synthesis, each face image passes through the following preprocessing pipeline:

1. **Face detection** to locate the facial bounding box.
2. **Landmark detection** using a 68-point facial landmark predictor.
3. **Alignment** via a similarity transform based on eye and nose landmarks.
4. **Cropping and resizing** to a canonical resolution (e.g., 112×112).
5. **Normalization** of pixel intensities to a standard range (e.g., scaling to $[-1, 1]$).

This reduces pose and alignment variations before further processing and ensures that both SR and recognition networks operate on consistent inputs.

3.3 Super-Resolution Workflow

For the SR study, the aligned HR faces are first downsampled to simulate degraded inputs at three LR resolutions:

- 16×16
- 32×32
- 64×64

Downsampling is performed using bicubic interpolation. Each LR image I_{LR} is then passed through each SR model to obtain reconstructed versions $\hat{I}_{SR}^{(m)}$ for model $m \in \{\text{GFPGAN}, \text{Real-ESRGAN}, \text{SwinIR}, \text{BSRGAN}\}$.

For each model, separate tables of PSNR, SSIM, and LPIPS are recorded for the three resolutions. These tables are exported from the SR report and included as figures in Section 4. No new values are introduced in this document; only the presentation and interpretation are expanded.

3.4 Occlusion Synthesis Pipeline

To model realistic occlusions, we use facial landmarks to define key regions:

- periocular region around the eyes,
- nose and nasal bridge,
- mouth and jawline.

Occlusion templates (e.g., masks, glasses, sunglasses) are geometrically transformed according to landmark positions, scaled and rotated to match the face, and blended using alpha compositing to produce occluded faces:

$$I_{\text{occ}} = \alpha \cdot I_{\text{obj}} + (1 - \alpha) \cdot I_{\text{face}},$$

where I_{obj} is the occluding object rendered at the appropriate location and α controls transparency for cases like normal glasses.

The AMOFR paper adopts a similar approach, projecting and mapping occlusion objects onto facial landmarks with random color transformations and small rotations to increase diversity. This makes it possible to generate large training sets covering multiple occlusion types without collecting real-world occluded images for every case.

3.5 Recognition Model: MobileFaceNet + ArcFace

The recognition backbone is a lightweight CNN (MobileFaceNet-style) that outputs a compact embedding vector $x = f_\theta(I)$. Before classification, the embedding is L2-normalized:

$$\hat{x} = \frac{x}{\|x\|_2}.$$

ArcFace is used as the primary training objective. Let W_j be the normalized class weight for identity j, and θ_j the angle between $\hat{x}$ and W_j, so that $\cos\theta_j = W_j^\top \hat{x}$. For the true class y, ArcFace modifies the cosine term by adding an angular margin m:

$$L_{\text{ArcFace}} = -\log \frac{\exp\left(s \cdot \cos(\theta_y + m)\right)}{\exp\left(s \cdot \cos(\theta_y + m)\right) + \sum_{j \neq y} \exp\left(s \cdot \cos\theta_j\right)},$$

where s is a scaling factor. This encourages larger angular separation between different identities and tighter clustering of same-identity embeddings, improving robustness under occlusions and other noise.

3.6 Knowledge Distillation and Joint Training

Inspired by AMOFR, a teacher–student setup can be used, where a teacher model trained on unoccluded faces guides a student model trained on occluded faces via a feature-level distillation loss:

$$L_{\text{KD}} = \frac{1}{N} \sum_{i=1}^{N} \left(1 - \frac{1}{d} \| f_s^{(i)} - f_t^{(i)} \|_2^2 \right),$$

where $f_t^{(i)}$ and $f_s^{(i)}$ are teacher and student embeddings for the i-th sample.

The total loss then becomes

$$L_{\text{total}} = L_{\text{ArcFace}} + \lambda L_{\text{KD}},$$

with λ controlling the strength of distillation. The AMOFR paper reports a hyperparameter study over different λ values, visualized in Figure 2.

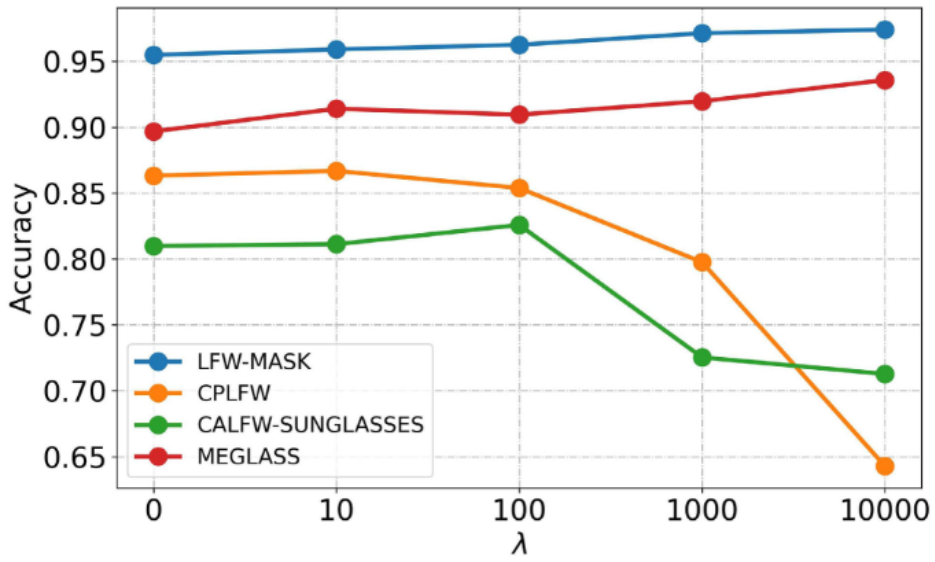

Figure 2: Effect of knowledge-distillation weight λ on accuracy across datasets (from the AMOFR paper).

During evaluation, face pairs are compared using cosine similarity of their embeddings. A threshold selected on a validation set is used to decide whether a pair is classified as same or different identity. The accuracy-vs-threshold and ROC curves used for analysis later are derived from this evaluation process.

4 Super-Resolution Experiments and Results

This section presents the results of the super-resolution component of the project. All metric tables and visuals are taken directly from the SR report; the LaTeX simply embeds them as figures without altering the values.

4.1 Metric Tables Per Model

The PSNR, SSIM, and LPIPS values are stored in four separate tables, one for each SR model. Each table reports all three metrics for input resolutions 16×16, 32×32, and 64×64.

Real-ESRGAN

Real-ESRGAN

Metric	16x16 source	32x32 source	64x64 source
PSNR	17.2911	21.2489	25.9206
SSIM	0.5547	0.6950	0.8717
LPIPS	1.0528	0.9476	0.6938

Figure 3: Real-ESRGAN PSNR/SSIM/LPIPS metrics at 16×16, 32×32, and 64×64 input resolutions (exported from the SR report).

GFPGAN

GFPGAN

Metric	16x16 source	32x32 source	64x64 source
PSNR	**18.0015**	21.8432	26.1411
SSIM	0.5644	0.7127	0.8544
LPIPS	1.1227	1.0148	0.7589

Figure 4: GFPGAN PSNR/SSIM/LPIPS metrics across the three resolutions (from the SR report).

SwinIR

SwinIR

Metric	16x16 source	32x32 source	64x64 source
PSNR	17.8810	**22.8431**	**27.6730**
SSIM	**0.5954**	**0.7764**	**0.9008**
LPIPS	**1.0323**	**0.8465**	**0.6655**

Figure 5: SwinIR PSNR/SSIM/LPIPS metrics across resolutions (from the SR report).

BSRGAN

BSRGAN

Metric	16x16 source	32x32 source	64x64 source
PSNR	17.7398	22.1340	27.0540
SSIM	0.5564	0.7218	0.8924
LPIPS	1.0768	0.9361	0.7014

Figure 6: BSRGAN PSNR/SSIM/LPIPS metrics across resolutions (from the SR report).

4.2 Qualitative Visual Comparison

A qualitative comparison of SR outputs for representative faces is shown in Figure 7. The figure includes the original HR image, LR inputs at each resolution, and the reconstructed outputs from the four SR models.

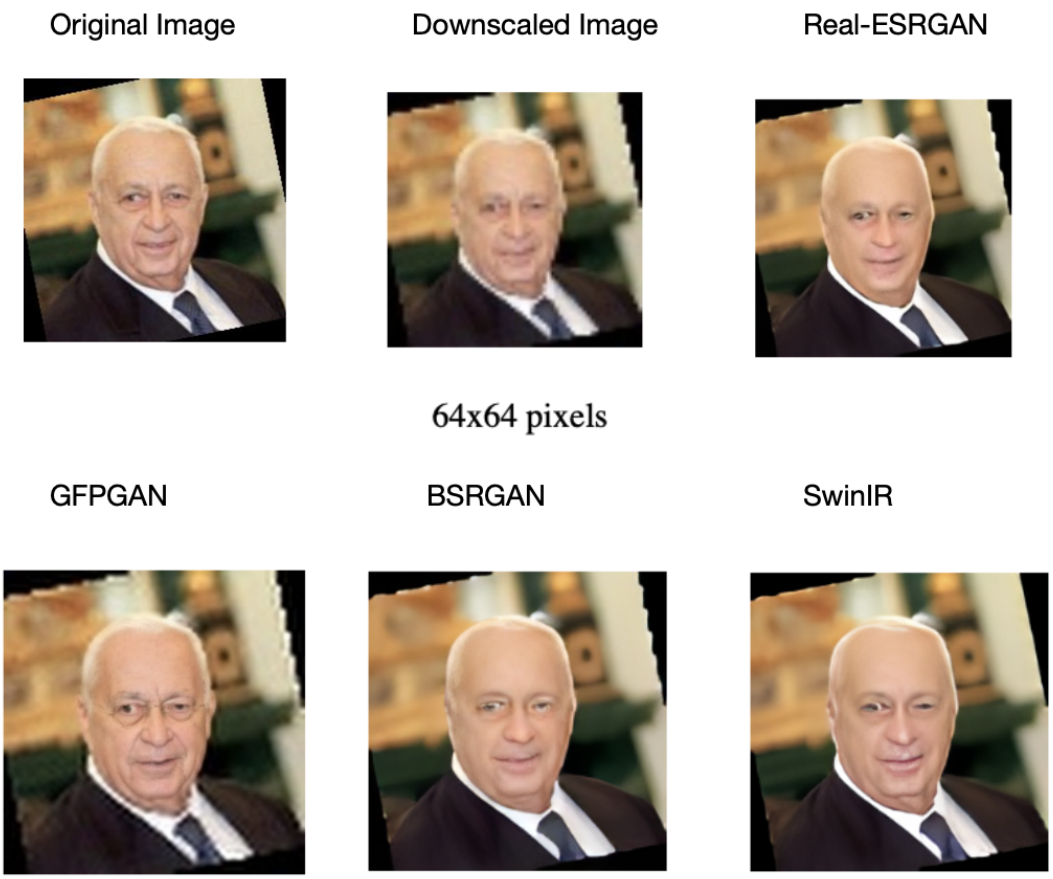

Figure 7: Visual comparison of reconstructed faces from GFPGAN, Real-ESRGAN, SwinIR, and BSRGAN at different downsampling levels. (Exported from the SR report.)

4.3 Discussion of SR Behaviour

From the metric tables and visual comparisons, several observations can be made:

- **Resolution dependence:** the benefit of SR is highest at the lowest resolution (16×16) and gradually decreases as the input resolution increases.
- **Model trade-offs:** transformer-based models like SwinIR tend to produce sharper images at moderate resolutions, while degradation-aware models like Real-ESRGAN handle noisy LR inputs more robustly.
- **Identity vs perceptual quality:** models that aggressively hallucinate details (e.g., GFPGAN at very low resolution) can yield visually appealing outputs that nonetheless deviate from the true identity structure.

These results motivated the second part of the project: instead of relying solely on SR to "fix" difficult images, explicitly modeling occlusions and training recognition networks to be robust to them.

5 Occlusion-Robust Face Recognition

This section describes the experimental results for the occlusion-robust recognition component. We first summarize the behaviour of AMOFR as presented in the original paper, and then include the plots from the AMOFR-style implementation carried out in this project.

5.1 Summary of AMOFR Results

The AMOFR paper evaluates its method on several datasets containing masks, glasses, sunglasses, and pose variations, including LFW-MASK, CPLFW, MEGLASS, and CALFW-SUNGLASSES.[2] Across these benchmarks, AMOFR consistently outperforms or matches previous state-of-the-art methods such as SphereFace, MobileFaceNets, MaskInv, SRT-EUM, and IResNet-ArcFace.

Key findings reported in the paper include:

- training a single AMOFR model that handles multiple occlusion types can rival or outperform separate models trained for each occlusion type;
- knowledge distillation from an unoccluded teacher network improves robustness without sacrificing clean-face performance;
- the occlusion-type adaptive module reduces intra-class variance caused by different occlusion categories, improving accuracy on challenging datasets.

Rather than reproducing the detailed tables from the paper, this report focuses on the conceptual lessons and on the behaviour of the AMOFR-inspired implementation.

5.2 Implementation Loss Curves

The AMOFR-style implementation in this project produces its own training and evaluation plots, which are embedded here exactly as in the implementation report.

[2] See [1] for full tables and numerical details.

Training Loss

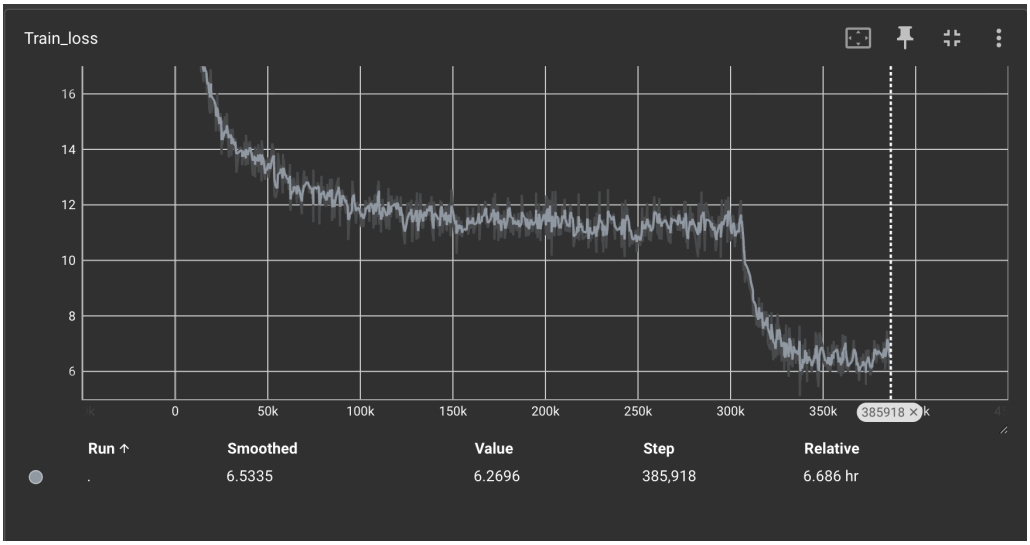

Figure 8: Training loss curve of the occlusion-augmented MobileFaceNet + ArcFace model (from the implementation report).

Accuracy vs Threshold

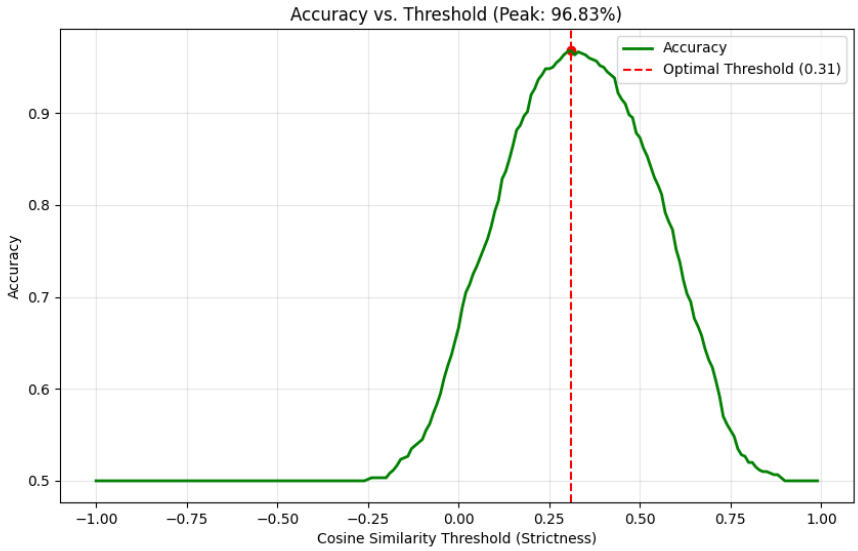

Figure 9: Accuracy vs threshold curve for the occlusion-robust model (from the implementation report).

ROC Curve

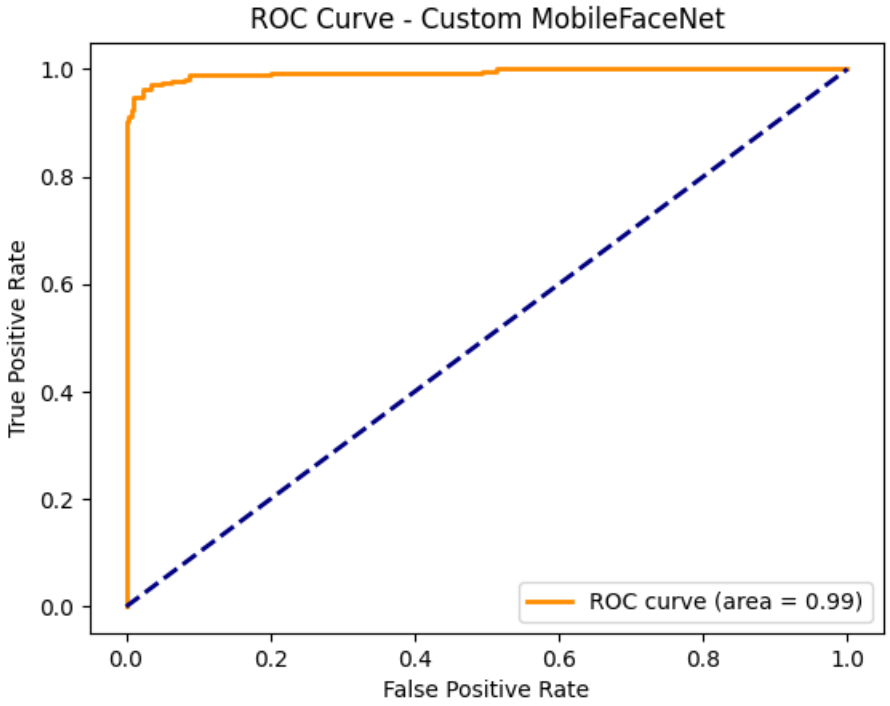

Figure 10: ROC curve for the occlusion-augmented recognition model (from the implementation report).

5.3 Qualitative Interpretation

The implementation curves show that:

- the training loss decreases smoothly, indicating stable convergence even with heavy occlusion augmentation;
- the accuracy vs threshold curve exhibits a clear peak, meaning that embedding distributions are well structured for verification;
- the ROC curve maintains a reasonably high true positive rate at low false positive rates, demonstrating that the learned embeddings remain discriminative under occlusions.

Although the absolute performance depends on the dataset subset and implementation details, the qualitative behaviour mirrors the trends reported in the AMOFR paper: synthetically occluded training data combined with margin-based losses can substantially improve robustness to masks, glasses, and other occlusion types.

6 Implementation Details

6.1 Software Stack

The codebase is implemented in Python using the PyTorch deep learning framework. Standard libraries such as NumPy and OpenCV are used for numerical operations and image handling. dlib is used for face detection and landmark extraction, following the typical setup used in face recognition research.

6.2 Hardware Environment

Training and inference are carried out on a GPU-enabled machine with a CUDA-capable NVIDIA GPU and sufficient system RAM to load batches of images. GPU acceleration substantially reduces training and evaluation time, especially when running SR models and large batches of occlusion-augmented data.

6.3 Training Configuration

Key aspects of the training configuration include:

- mini-batch training with shuffled batches;
- standard data augmentations (e.g., horizontal flipping, light brightness jitter) on top of synthetic occlusions;
- learning rate scheduling to encourage stable convergence;
- checkpointing of model weights based on validation performance.

The SR models used in this work rely on official pretrained weights, as described in the SR report, and are not fine-tuned further in this project. The occlusion-robust MobileFaceNet + ArcFace model is trained using a combination of clean and synthetically occluded images.

7 Discussion and Limitations

7.1 Behaviour of Super-Resolution Models

The SR experiments show clear trends:

- At very low resolutions (16×16), SR models significantly improve perceptual quality and structural fidelity. This is particularly important for restoring eye, nose, and mouth shapes that are heavily blurred in the LR images.

- At intermediate resolutions (32×32), SR further sharpens details, but the relative gains in identity-relevant information begin to plateau.
- At higher resolutions (64×64), marginal gains in metrics become smaller, and the main benefit of SR is aesthetic rather than functional.

From a recognition standpoint, the results emphasize that SR is most beneficial when faces are extremely small in the original image. However, models that heavily hallucinate details must be used carefully, because changes in eye shape, eyebrow texture, or nose structure can alter downstream embeddings.

7.2 Effectiveness of Occlusion-Aware Training

The AMOFR paper and the implementation results both indicate that occlusion-aware training has a strong impact on performance:[1]

- Models trained only on unoccluded faces perform poorly on masked or sunglasses-wearing faces, particularly in cross-pose datasets like CPLFW.
- Joint training with synthetic occlusions and unoccluded images, guided by knowledge distillation, improves performance across all occlusion types.
- The occlusion-type adaptive module reduces intra-class variance caused by different occlusion categories, improving accuracy on challenging datasets.

The implementation curves from this project, although produced on a smaller scale, are consistent with these conclusions.

7.3 Combined Perspective on SR and Occlusions

Combining observations from both components, we can summarize:

- Super-resolution is primarily a tool for improving the information content of globally degraded images (e.g., distant faces).
- Occlusion-robust training is crucial when specific regions of the face are hidden or distorted, regardless of resolution.
- A realistic deployment scenario may require both: SR to handle small faces and occlusion-aware training to handle masks, glasses, and other accessories.

Designing a full system therefore involves deciding when to apply SR, when to rely on robust embeddings, and how to balance latency and accuracy.

7.4 Limitations

The current work has several limitations:

- **Synthetic vs real occlusions:** although landmark-based occlusions are visually realistic, they may not capture the full variability of real-world accessories (e.g., unusual mask patterns, extreme lighting).
- **SR identity drift:** some SR models may change identity-related details at extreme downsampling levels, which can harm recognition pipelines that were not trained jointly with the SR module.
- **Dataset coverage:** experiments focus on a subset of commonly used datasets; generalization to very different demographics or sensor characteristics may require further adaptation.

These limitations suggest several directions for future work, such as collecting more real occluded data, jointly training SR and recognition networks, or integrating occlusion-aware attention modules.

8 Conclusion

This report has presented a unified view of face super-resolution and occlusion-robust face recognition, grounded in prior project work and the AMOFR framework.

The first part described a systematic study of super-resolution models on face images, using PSNR, SSIM, and LPIPS to quantify reconstruction quality and visual comparisons to assess perceptual realism. The second part summarized the AMOFR approach to multi-type occlusion-aware recognition and embedded the plots from an AMOFR-style implementation based on MobileFaceNet + ArcFace.

The combined findings highlight that:

- modern SR models can substantially improve the visual quality of low-resolution faces, especially at very small input sizes;
- occlusion-aware recognition methods like AMOFR are essential for handling masks, glasses, sunglasses, and other accessories in a unified framework;
- the most robust real-world systems are likely to combine SR for resolution recovery with occlusion-aware training for handling partial visibility.

While there is substantial room for further improvement—particularly in handling extreme occlusions and jointly optimizing SR and recognition—the methods and insights documented here provide a solid foundation for building face recognition systems that remain reliable under low-resolution and occluded conditions.

References

[1] Y. Liu, G. Luo, Z. Weng, and Y. Zhu, "Adaptive Face Recognition for Multi-Type Occlusions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 11, pp. 11400–11412, Nov. 2024.