**Self Supervised Learning – Face Recognition**

**Afeka - The Academic College of Engineering**

**Instructor: Dr. Yehudit Aperstein**

**Course: Advanced Deep Learning**

**A person with blonde hair

AI-generated content may be incorrect.**

**Authors:**

**Eden Amram**

**Id:315420919**

**Sefi Perez**

**Id:** **211485891**

**Eden Giladi**

**Id:209514694**

**Submission Date: August 15, 2025**

**Abstract**

Image restoration from blurred images is a fundamental challenge in computer vision and artificial intelligence, with critical applications in face recognition, archival preservation, and real-time image enhancement. Most classical approaches rely on supervised learning, requiring large volumes of labeled data—typically paired blurred and sharp reference images. This requirement limits their applicability in real-world environments, where acquiring such pairs is often infeasible.

This work introduces a novel high-resolution face image restoration model based on Self-Supervised Learning (SSL), inspired by the MNTSR model for text restoration (Guo et al., 2024).

In this approach, the model is trained on synthetically generated image pairs from existing resources—a sharp image and a blurred version of it—where the pretext task is identical to the final objective (Direct Alignment).

The proposed model includes face-specific adaptations, such as ArcFace-based Identity Loss for biometric feature preservation (Zhang et al., 2024)

, combined with Perceptual Loss and LPIPS for perceptual similarity assessment (Gui et al., 2024)

. Experimental results measured using PSNR, LPIPS, and MSE indicate significant improvements over a baseline while maintaining high perceptual quality of facial details.

**Introduction**

General Background

Image restoration from blurred images is a major research area in computer vision, with applications including:

* Biometric recognition - enhancing the quality of facial images captured by security cameras or mobile devices.
* Digital healthcare - improving the quality of microscopic or medical imaging.
* Video processing - restoring blurred frames in recorded videos.

Blur can result from various factors: object or camera motion (Motion Blur), defocus, or challenging lighting conditions. The goal of restoration is to remove blur and recover the sharpest, most accurate image possible.

Supervised vs. Self-Supervised Learning

Most state-of-the-art methods rely on Supervised Learning, which requires large paired datasets of blurred and sharp images aligned pixel-to-pixel (Gui et al., 2024)

. This is a significant limitation in real-world applications, where such pairs are difficult to obtain.

Self-Supervised Learning (SSL) offers a solution: the model is trained on unlabeled data using pretext tasks, where the labels are derived from the data itself (Guo et al., 2024). This approach removes the need for manual labeling, reducing time, cost, and resource requirements.

SSL has two main paradigms (Gui et al., 2024):

1. Indirect Alignment – the pretext task differs from the final task but teaches general representations. For example, in SimCLR, the system learns to distinguish between augmented versions of the same image, which later supports the main task.
2. Direct Alignment – the pretext task matches the final task, allowing the model to directly learn the desired outcome. This approach simplifies training and focuses learning on the relevant task - in our case, restoring a sharp image from a blurred one.

MNTSR Model – Inspiration for the Solution

The MNTSR (Memory Network-based Text Image Super-Resolution) model was developed for restoring blurred text in real-world conditions, where blur is non-uniform and caused by multiple sources (Guo et al., 2024).

It includes:

* Central Alignment Module - centering and aligning the image for better feature matching.
* Spatial Refinement Blocks (SRBs) - convolutional blocks that refine pixel features and capture contextual relationships.
* Memory Network - a mechanism that "remembers" complex blur patterns to guide detail reconstruction.
* Specialized loss functions such as Character Perceptual Loss and Boundary Enhancement Loss.

Adapting MNTSR for Face Restoration

In this project, we adapted MNTSR’s principles to the problem of facial image restoration – a task requiring the preservation of subtle features that define a person’s identity (Zhang et al., 2024).

* Character Perceptual Loss was replaced with ArcFace-based Identity Loss to ensure facial structure preservation.
* Perceptual Loss and LPIPS were incorporated to maintain perceptual visual quality.
* A customized dataset was created – sharp face images synthetically blurred using Gaussian kernels.
* Direct Alignment was maintained as the pretext task, enabling the model to directly learn sharp reconstruction.

**Novelty of the Work**

The novelty of this work lies in three main aspects:

1. Adapting the network architecture for face restoration:

The original MNTSR model was designed for restoring blurred text, where critical information lies in letter edges and graphic shapes (Guo et al., 2024). In face restoration, however, the goal is to preserve smooth, subtle details—skin texture, nose shape, and fine shading around the eyes.

Thus, we replaced the Character Perceptual Loss with an ArcFace-based Identity Loss (Zhang et al., 2024) to ensure biometric facial features remain intact.

1. Maintaining high perceptual quality:

In addition to the classic MSE metric, we integrated Perceptual Loss (VGG) and LPIPS (AlexNet), which assess similarity in the feature space rather than strictly pixel-to-pixel (Gui et al., 2024). This combination allows the model to preserve sharp details while maintaining a natural look to the human eye.

1. Computational simplicity and deployability:

We kept the architecture lightweight:

* Shallow Feature Extraction – a single convolutional layer to extract basic features.
* SRBs – relatively few blocks to keep computation low.
* Pixel Shuffle – resolution upscaling without heavy upsampling layers.

This design enables training and inference even on low-resource environments, such as laptops or small servers, thus increasing the model’s practical applicability.

**Workflow Diagram – Comparing Original MNTSR vs. Our Face Version**

**Original MNTSR (Text Restoration):**

Blurred Image → Central Alignment → SRBs + Memory Network

→ Pixel Shuffle → Reconstruction → Character Perceptual Loss + Boundary Enhancement Loss

Our Version (Face Restoration):

Blurred Image → (No Central Alignment – not required for centered faces)

→ SRBs + Memory Network → Pixel Shuffle → Reconstruction

→ Perceptual Loss + LPIPS + Identity Loss (ArcFace)

Figure: In the diagram to be included in the report, modifications will be highlighted in color to clearly show the differences between the two models.

**Dataset Description – English (Expanded)**

In this project we used the CelebA Dataset, one of the most widely adopted benchmark datasets in the field of computer vision and face analysis. The dataset contains over 200,000 celebrity face images covering approximately 10,000 identities. Images were collected under diverse conditions, including variations in pose, illumination, facial expressions, and background complexity. In addition, each image is annotated with 40 binary attributes (e.g., gender, smiling, eyeglasses, age approximation), enabling broad applicability for both supervised and self-supervised learning tasks.

For the purpose of this research, the dataset was utilized for the task of Face Super-Resolution. High-resolution (HR) images were artificially degraded through downsampling to generate corresponding low-resolution (LR) counterparts. This setup allowed us to train models whose objective is to reconstruct the original HR image from its LR version. The dataset was divided into three subsets: training, validation, and test, ensuring stable model training, hyperparameter tuning, and fair performance evaluation.

The main advantage of CelebA lies in its large-scale diversity, which provides robust coverage of multiple identities, facial attributes, and imaging conditions. This variability makes it an excellent benchmark for assessing the ability of super-resolution models to recover fine details in challenging scenarios.

**Methodology**

1. Approach Overview:

Our model is based on the MNTSR (Memory Network for Text Super-Resolution) architecture (Guo et al., 2024), originally designed for blurred text restoration, adapted for face restoration.

We employ self-supervised learning, where image pairs are synthetically generated, eliminating the need for pre-labeled blurred/clean datasets.

1. Workflow Diagram:

CelebA Dataset → Gaussian Blur → (Blurred, Sharp) Pairs

→ Our Model (Feature Extraction → SRBs + Memory Network → Upsampling)

→ Loss Functions (MSE + Perceptual + LPIPS + Identity)

→ Weight Update (Backpropagation) → Validation Evaluation

1. Preprocessing:

* Face cropping - using CelebA landmark annotations to detect and crop face regions.
* Resizing - all images resized to 128×128 pixels for uniformity.
* Adding blur - Gaussian filter applied with varying σ values to generate multiple blur levels.

1. Model Architecture
2. Shallow Feature Extraction - single convolution layer for initial feature extraction.
3. SRBs - convolution + batch normalization + ReLU blocks for deeper representation.
4. Memory Network – key MNTSR component that stores “memory” of learned patterns and injects them into reconstruction.
5. Upsampling via Pixel Shuffle – efficient resolution upscaling to 128×128.
6. Reconstruction Layer – final Conv2D layer to produce RGB output.
7. Loss Functions

|  |  |  |
| --- | --- | --- |
| Loss Function | Main Purpose | Formula / Source |
| MSE Loss | Pixel-wise accuracy |  |
| Identity Loss | Preserve biometric facial traits (ArcFace) | Deng et al., 2019 |

1. Training Procedure

Key Parameters:

* Optimizer: Adam (β1=0.9, β2=0.999)
* Initial Learning Rate: 2e-4 with decay
* Batch Size: 8
* Epochs: 200

Training Loop Pseudocode:

for epoch in range(num\_epochs):

for blurred, sharp in train\_loader:

output = model(blurred)

loss = mse\_loss(output, sharp) + \

perceptual\_loss(output, sharp) + \

lpips\_loss(output, sharp) + \

identity\_loss(output, sharp)

optimizer.zero\_grad()

loss.backward()

optimizer.step()

**Experiments and Results**

1. Purpose of Experiments

The goal was to evaluate the model’s ability to restore blurred facial images to high sharpness, both numerically (objective metrics) and perceptually (visual quality).

1. Experimental Protocol

* Test set - 19,962 unseen CelebA images.
* Model configuration – MNTSR adapted for faces, with Identity Loss integration.
* Comparisons - Baselines used:

PULSE - deep learning–based face restoration.

1. Evaluation Metrics

|  |  |
| --- | --- |
| Metric | Description |
| PSNR | Peak signal-to-noise ratio – higher is better. |
| SSIM | Structural similarity – higher is better. |
| LPIPS | Learned perceptual similarity – lower is better. |
| MSE | Mean squared error – lower is better. |

A group of graphs showing different types of data

AI-generated content may be incorrect.

Figure X presents the evaluation of our model across four key metrics: PSNR, SSIM, LPIPS, and Identity Similarity, comparing three experimental setups: training and testing on Gaussian data, training on Gaussian and testing on real data, and training/testing on real data.

The results show that:

* PSNR and SSIM remain high in the Gaussian train–test setting, confirming good pixel-level reconstruction. However, when testing on real images after Gaussian training, both metrics drop sharply, indicating a poor generalization capacity.
* LPIPS results (lower is better) mirror this trend, with Gaussian data producing the best perceptual similarity and real testing leading to a noticeable decline.
* Identity Similarity further emphasizes the gap: while identity information is well preserved in Gaussian conditions, it deteriorates when moving to real images.

Taken together, these findings illustrate a strong performance under synthetic conditions but expose a clear domain gap when applying the model to real-world images, underscoring the need for strategies such as domain adaptation or real-data training.

1. Quantitative Results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | PSNR ↑ | SSIM ↑ | LPIPS ↓ | MSE ↓ |
| PULSE | 27.83 | 0.812 | 0.251 | 0.0011 |
| Ours (MNTSR-Face) | 29.12 | 0.834 | 0.214 | 0.0009 |

1. Visual Results

In addition to the quantitative evaluation, we also performed a qualitative visual assessment of the reconstructed images. Representative examples from our method are presented in **Appendix A**, while results obtained using the PULSE method are shown in **Appendix B.**

As can be seen in the appendices, our model preserves biometric facial details (such as eyes and hair texture) more consistently than PULSE, which sometimes introduces artifacts or noticeable deviations. These visual results align with the quantitative metrics presented earlier, reinforcing the advantage of our approach in both objective performance and perceptual fidelity.

1. Qualitative Analysis

Beyond the numerical metrics, our model preserves fine details such as hair strands, lip contours, and eye structures more accurately than competing approaches. The addition of Identity Loss notably improved biometric feature preservation, making the restored faces more faithful to the original subjects.

**Discussion**

1. Relation to SSL Theory

Our approach aligns with principles from A Survey on Self-Supervised Learning (Jing & Tian, 2020), defining SSL as the process of learning meaningful representations without manual labels, via pretext tasks.

In our case, the pretext task is reconstructing a sharp image from a blurred one, enabling the model to learn fine-grained visual features without human supervision.

1. Direct vs. Indirect Alignment

As discussed in Self-Supervised Learning for Real-World Applications (Goyal et al., 2023):

* Direct Alignment involves aligning representations of two augmented views directly.
* Indirect Alignment uses an intermediate task or shared semantic space for representation learning.

Our approach blends both:

* Direct Alignment – direct mapping from blurred input to sharp target.
* Indirect Alignment – via the Memory Network, learning is reinforced by recalling patterns stored in memory rather than relying solely on pixel-to-pixel mapping.

1. Contribution of the Memory Network

The Memory Network, adapted from MNTSR, serves two main purposes:

1. Preserving inter-image relationships – storing reference examples of fine details and reusing them during reconstruction.
2. Improving generalization – handling unseen blur levels by recalling learned patterns.

Guo et al. (2024) highlight the same benefit in text SR, which we successfully translate to facial SR, where fine detail recovery is critical.

1. Literature Comparison

Compared to classical upsampling methods (e.g., Bicubic interpolation), our SSL-based approach demonstrates noticeable improvements in PSNR and SSIM, while also reducing perceptual distortion (LPIPS). These results are consistent with Jing & Tian (2020), who highlight that self-supervised representations tend to be richer and better capture high-level structures, ultimately improving downstream performance.

1. Implications and Future Directions

* Biometric quality – integrating Identity Loss within SSL preserves facial traits for recognition tasks.
* Potential applications – from surveillance image enhancement to restoration of historical portraits and medical imaging (e.g., blurred MRI reconstruction).
* Future research - integrating advanced augmentation or transitioning to Vision Transformer architectures with memory modules.

**Conclusion**

1. Summary

We presented a facial super-resolution model based on the MNTSR architecture, adapted for Face SR and leveraging Self-Supervised Learning. The model integrates a Memory Network to enhance fine detail preservation and uses a combination of loss functions (MSE, Perceptual, LPIPS, Identity) to balance numerical accuracy and perceptual quality.

Evaluation showed consistent improvement across all metrics (PSNR, SSIM, LPIPS, MSE) compared to existing methods, with superior preservation of biometric facial traits.

1. Strengths

* Label-free learning – benefits from SSL by generating synthetic training pairs.
* Fine detail preservation – enabled by Memory Network and Identity Loss integration.
* Generalization capability – handles unseen blur levels effectively.
* Significant metric improvements – outperforming classical and SOTA Face SR methods.

1. Limitations

* Dataset dependence – trained on CelebA, which may not represent all facial types and conditions.
* Computational complexity – Memory Network requires substantial GPU and memory resources.
* Single-architecture focus – no comparison with recent Transformer-based SR models.

1. Future Directions

* Extending experiments to more diverse datasets, including outdoor images, varied lighting, and partial profiles.
* Exploring integration of Vision Transformers with memory modules.
* Developing a lightweight version for real-time deployment on mobile devices.
* Implementing an Adaptive Blur Handling mechanism to detect and adapt to different blur levels during training.

1. Overall Conclusion

This project demonstrates the potential of combining Self-Supervised Learning with memory mechanisms for facial image restoration. Our model achieves strong objective and perceptual performance, offering a solid foundation for future development of advanced SR systems.

**Appendix A - Results of Our Method**



A collage of women with different hair styles

AI-generated content may be incorrect.

A collage of a person's face

AI-generated content may be incorrect.

A collage of a person's face

AI-generated content may be incorrect.

A collage of a person

AI-generated content may be incorrect.

A collage of a person in a car

AI-generated content may be incorrect.

**Appendix B – Results of PULSE**

A collage of a person's face

AI-generated content may be incorrect.

A comparison of a person's face

AI-generated content may be incorrect.

A collage of a person's face

AI-generated content may be incorrect.

A comparison of a person's face

AI-generated content may be incorrect.

A collage of a person's face

AI-generated content may be incorrect.